Roberto Casarin* and Antonio Peruzzi

A Dynamic Latent-Space Model for Asset Clustering

https://doi.org/10.1515/snde-2022-0111
Received November 30, 2022; accepted October 16, 2023; published online November 23, 2023

Abstract: Periods of financial turmoil are not only characterized by higher correlation across assets but also by modifications in their overall clustering structure. In this work, we develop a dynamic Latent-Space mixture model for capturing changes in the clustering structure of financial assets at a fine scale. Through this model, we are able to project stocks onto a lower dimensional manifold and detect the presence of clusters. The infinite-mixture assumption ensures tractability in inference and accommodates cases in which the number of clusters is large. The Bayesian framework we rely on accounts for uncertainty in the parameters' space and allows for the inclusion of prior knowledge. After having tested our model's effectiveness and inference on a suitable synthetic dataset, we apply the model to the cross-correlation series of two reference stock indices. Our model correctly captures the presence of time-varying asset clustering. Moreover, we notice how assets' latent coordinates may be related to relevant financial factors such as market capitalization and volatility. Finally, we find further evidence that the number of clusters seems to soar in periods of financial distress.

Keywords: latent space models; Bayesian inference; financial risk

JEL Classification: C11; C32; G1

1 Introduction

That average correlation across stocks surges during periods of financial distress is something the finance literature is well aware of (e.g. see Corsetti et al. 2001; Corsetti, Pericoli, and Sbracia 2005; Forbes and Rigobon 2002). See also Corsetti, Pericoli, and Sbracia (2011) for a review of stylized facts on contagion and interdependence. Drawing on these seminal papers, a large body of literature has investigated contagion effects (e.g. see Casarin, Sartore, and Tronzano 2018; Münnix et al. 2012; Preis et al. 2012; Zheng et al. 2012). However, there is also evidence that not only cross-correlation increases on average but also that the clustering structure of the correlation matrix varies dynamically (Ahelegbey, Carvalho, and Kolaczyk 2020; Kocheturov, Batsyn, and Pardalos 2014; Nie 2017). The German and the US stock markets make no exception in this respect. Figure 1 reports the estimated densities of the yearly cross-correlation of DAX40 (Panel A) and S&P100 (Panel B) constituents through time. We observe soaring average cross-correlation in periods of distress: i.e. the 2008–2009 financial crisis, the European debt crisis in 2011, the stock-market sell-off in 2015, and the Covid-19 outbreak in 2020. Bi-modality, which can be considered a symptom of clustering, emerged during some of the aforementioned years. This is particularly evident from the DAX40 correlation structure.

We deem that the possibility of providing an in-depth analysis of assets correlation structure and its dynamics is of major relevance and interest to researchers and investors. Many diversification strategies fail during

***Corresponding author: Roberto Casarin**, Department of Economics, Venice Centre in Economic and Risk Analytics for Public Policies (VERA), Ca' Foscari University of Venice, Venice, Italy, E-mail: r.casarin@unive.it. https://orcid.org/0000-0003-1746-9190
Antonio Peruzzi, Department of Economics, Ca' Foscari University of Venice, Venice, Italy, E-mail: antonio.peruzzi@unive.it. https://orcid.org/0000-0001-8865-943X

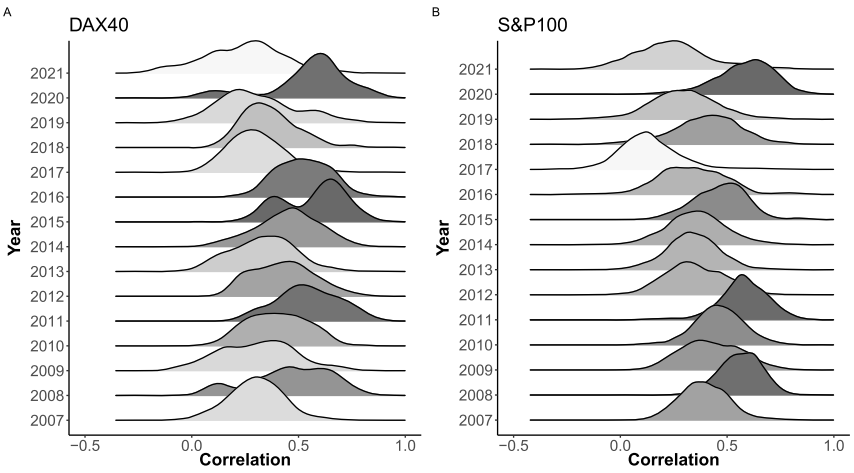

Figure 1: Cross-correlation in DAX40 and S&P100. Panel A: densities of DAX40 constituents cross-correlation distribution through time. Panel B: densities of S&P100 constituents cross-correlation distribution through time. We notice how average cross-correlation increases during years of financial turmoil and how bi-modality – which can be considered a symptom of clustering – emerges, especially in DAX40, in those same years.

periods of financial distress as portfolio components get more and more correlated. Thus, analyzing the collective behavior of assets is a crucial step in asset allocation and risk management. Some explanations regarding assets' collective behavior have been proposed by Onnela et al. (2003) and Khandani and Lo (2007). The application of clustering methods to financial assets became popular after the early works of Farrell (1974), King (1966), and Arnott (1980). Our model allows for analyzing assets' proximity and clustering structure of assets' correlation matrices.

The large dimension of assets' correlation matrices calls for dimensionality reduction techniques, which are growing in popularity due to the increased availability of large datasets. Within these techniques, those based on low-rank representations and Bayesian inference procedures revealed quite effective in many classical econometric problems (Baştürk, Hoogerheide, and van Dijk 2017) and have been successfully extended to high-dimensional models (e.g. see Billio, Casarin, and Iacopini 2022a; Billio et al. 2022b; Hoff 2021).

In this work, we apply a low-rank representation framework based on Latent Space (LS) models, which has been introduced by Hoff, Raftery, and Handcock (2002). LS models allow for representing multiple-dimension data, such as matrices, networks, and tensors, on a lower d-dimensional latent space. See Raftery (2017) for a review. Several modifications to the basic logistic-link LS model have been implemented. Friel et al. (2016) provide a dynamic LS model suitable for bipartite-network data. Sewell and Chen (2016) introduced a dynamic LS model for weighted networks. Smith, Asta, and Calder (2019) explored the geometry of the latent space in relation to the network structure. One of the first extensions by Handcock, Raftery, and Tantrum (2007) regarded the introduction of a finite mixture on the set of latent coordinates. This model has been extended by D'Angelo (2018) to account for an infinite number of clusters using Dirichlet Process priors.

In this work, we develop a dynamic extension of the infinite-mixture LS model. The model allows for clustering in the correlation among asset log returns with an unknown number of clusters and for time variations in the clustering effects. We follow a Bayesian nonparametric approach to inference, contributing to the literature on Bayesian nonparametrics based on Dirichlet process priors and their extensions. These techniques find application in econometrics (e.g. see Bassetti, Casarin, and Leisen 2014; Bassetti, Casarin, and Ravazzolo 2018; Billio, Casarin, and Rossini 2019; Fisher and Jensen 2022a; Griffin and Kalli 2018; Griffin and Steel 2011; Hirano 2002; Jensen and Maheu 2010; Nieto-Barajas and Quintana 2016; Taddy and Kottas 2009) and in many other fields, such as biostatistics (Do, Müller, and Tang 2005), biology (Arbel, Mengersen, and Rousseau 2016), medicine (Xu et al. 2016), and neuroimaging (Zhang et al. 2016). See Hjort et al. (2010) for an introduction to Bayesian nonparametrics and (Müller and Mitra 2013) for a review of models and applications. Our model is well suited for

analyzing the temporal evolution of the location of the assets in the latent space (latent positions), their clustering effects, and the number of clusters.

Ahelegbey, Carvalho, and Kolaczyk (2020) propose one of the few applications of LS models for financial data analysis. The authors combine LS models and covariance-structure learning algorithms to extract asset latent positions from a network adjacency matrix. They extract the networks sequentially by a rolling estimation of a graphical VAR model and then estimate the LS clustering index on each network to measure contagion effects during periods of financial distress. Our LS model differs from the one proposed by Ahelegbey, Carvalho, and Kolaczyk (2020) as our model is dynamic, and our focus is on detecting the number of asset clusters over time. Moreover, we investigate the dependence between the number of clusters and some relevant market features, providing new empirical evidence.

The structure of the work now follows. Section 2 introduces the proposed model and describes some theoretical properties. Section 3 provides a Bayesian inference procedure and proves its efficiency through some simulation experiments. Section 4 presents the application to the constituents of DAX40 and S&P100 stock indices. Section 5 concludes the work.

2 A Dynamic Latent-Space Infinite Mixture Model

2.1 A Latent-Space Model

Let $\mathbf{C}_t$ $t = 1, 2, \dots, T$ be a sequence of N-dimensional correlation matrices. Each matrix is estimated in one of T non-overlapping time intervals on a panel of N log-return series $\mathbf{Y}_t = \{\boldsymbol{Y}_{1t}, \boldsymbol{Y}_{2t}, \dots, \boldsymbol{Y}_{Nt}\}$. Let w_{ijt}, for $i = 1, \dots, N$ and $j < i$, be the collection of M monotone transformations of the unique cross-correlations elements $\mathbf{C}_{ijt}$ of $\mathbf{C}_t$. We assume $w_{ijt} \overset{ind}{\sim} \mathcal{N}(\nu_{ijt}, \gamma_t^2)$, where $\mathcal{N}(\mu, \sigma^2)$ denotes a univariate Normal distribution described by a location parameter μ and scale parameter σ^2. We further assume that the location parameter for any pair of assets is decreasing for higher values of the squared distance of the assets on the latent space. We model such a relationship as follows:

$$\nu_{ijt} = \alpha_t - ||\mathbf{x}_{it} - \mathbf{x}_{jt}||^2, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean distance and $\mathbf{x}_{it} \in \mathbb{R}^{d_x}$ is a d_x-dimensional vector of latent coordinates. We assume a law of motion for the intercept parameter α_t:

$$\alpha_t = \alpha_{t-1} + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2), \tag{2}$$

with $\alpha_1 \sim \mathcal{N}\left(0, \sigma_{\alpha_0}^2\right)$ and $\sigma_{\alpha_0}^2 \sim \mathcal{IG}(a_{\alpha_0}, b_{\alpha_0})$, where $\mathcal{IG}(a, b)$ denotes an inverse-Gamma distribution with shape parameter a and scale parameter b.

We further assume that latent positions get drawn from an infinite mixture of normal distributions:

$$\mathbf{x}_{it} \sim \sum_{k_{it}=1}^{\infty} \lambda_{k_{it}} \mathcal{N}_{d_x}\left(\boldsymbol{\mu}_{k_{it}}, \sigma_{k_{it}}^2 I_{d_x}\right), \tag{3}$$

where $\mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$ denotes a d-variate Normal distribution with location $\boldsymbol{\mu}$ and scale Σ. Such a model allows not only for the estimation of the latent coordinates of assets according to a cluster structure but also it allows for an estimation of the number of clusters in which the latent positions can be partitioned.

2.2 Properties of Our Latent-Space Model

We present some relevant properties of a parametric LS model with an identity link function. In addition, we relate these properties with some characteristic features of weighted networks, such as the first and second moments of the nodal strengths.

Assumption 1. Given an undirected weighted network, $\mathcal{G}_t = (V, E_t)$, having vertex set $V \subset \mathbb{N}$ and weighted edge sets $E_t \subset V \times V$ and with characteristic weights w_{ijt}, we assume the existence of a sequence of d_x-dimensional latent variables $\{\mathbf{X}_{1t}, \dots, \mathbf{X}_{Nt}\}$ with $\mathbf{X}_{it} \in \mathcal{X} \subset \mathcal{R}^{d_x}$ for each node $i \in V$. The set $\mathcal{X}_t = \{\mathbf{x}_{1t}, \dots, \mathbf{x}_{it}, \dots, \mathbf{x}_{Nt}\}$ consists of i.i.d. realizations of these latent random variables, where each $\mathbf{x}_{it}$ is distributed according to $\pi(\cdot)$, a given probability measure.

Assumption 2. We assume conditional independence between any two edges given the latent variables. Hence, $\forall (j,i) \in E_t$, w_{ijt} is a normal random variable with mean parameter $\nu(\mathbf{x}_{it}, \mathbf{x}_{jt})$ and variance γ_t^2, that is $w_{ijt} \sim \mathcal{N}\left(\nu(\mathbf{x}_{it}, \mathbf{x}_{jt}), \gamma_t^2\right)$.

Under Assumptions 1 and 2, our LS model is an LVM model, as defined in Rastelli, Friel, and Raftery (2016), with an identity link. Moreover, we assume that our set of latent variables is jointly normally distributed. The normal distribution assumption can be replaced by the assumption of an infinite mixture of normals. Some of the properties presented later in this paper can be easily extended.

Assumption 3. The realized latent variables belonging to $\mathcal{X}_t$ are points in the Euclidean d-dimensional space, for a fixed d_x, and they are normally distributed:

$$p(\mathcal{X}_t) = \prod_{i=1}^{N} f_{d_x}\left(\mathbf{x}_{it}; \underline{0}, \sigma_t^2 I_{d_x}\right) = \prod_{i=1}^{N} \left(2\pi\sigma_t^2\right)^{-\frac{d_x}{2}} e^{-\frac{1}{2\sigma_t^2}\mathbf{x}_{it}'\mathbf{x}_{it}}.$$

We specify in more detail the form of the intensity parameter $\nu(\mathbf{x}_{it}, \mathbf{x}_{jt})$.

Assumption 4. Given the intercept parameter α_t, we assume:

$$\nu(\mathbf{x}_{it}, \mathbf{x}_{jt}) = \alpha_t - ||\mathbf{x}_{it} - \mathbf{x}_{jt}||^2.$$

Graph theory provides a useful approach to the analysis of the connectivity level in financial markets. See Bollobás (1998) and Diestel (2017) for an introduction to graph theory. In a network representation framework of correlations, the nodes of the network are the assets, the node labels are denoted by the index $i = 1, \dots, N$, and the weights of the edges among nodes are the transformed correlations w_{ijt} (see Bonanno et al. 2004; Namaki et al. 2011). There are many measures available for network connectivity (e.g. see Boccaletti et al. 2006; Newman 2018). The nodal strength is a local measure used as a first step in analyzing the connectivity of a network (Barrat et al. 2004). Thus, deriving the properties of such a measure is crucial for understanding the properties of a random graph model. In the following, we define the nodal strength s_{it}.

Definition 1. (Nodal Strength). Given the set of weights w_{ijt} we define the nodal strength s_{it} of i as the sum of all the weights w_{ijt} of i, that is:

$$s_{it} = \sum_{j \neq i} w_{ijt}. \tag{4}$$

The expected strength of a node in the weighted network coveys information on the local connectivity of such a network. A simple global connectivity measure can be obtained as the average of the local measures, as in the following definition.

Definition 2. (Average Nodal Strength). The network's average strength is the average of all individual strengths s_{it}:

$$\bar{s}_t = \frac{1}{N} \sum_{j \neq i} \sum w_{ijt}. \tag{5}$$

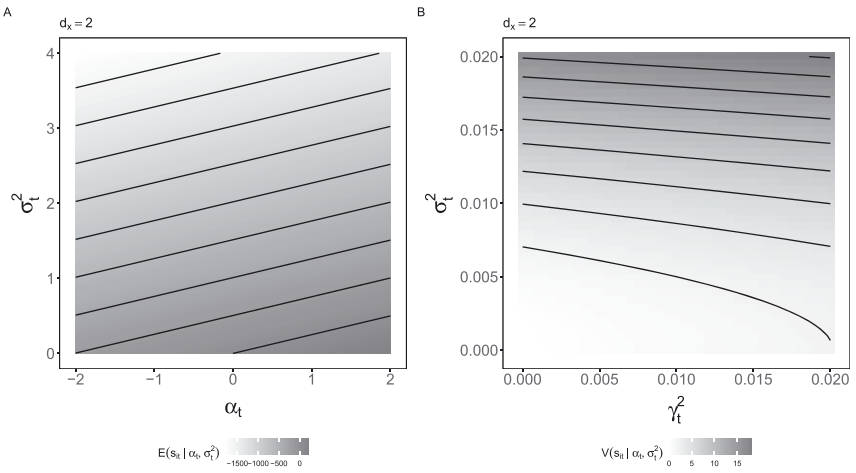

Figure 2: Strength properties. Panel A reports the expected strength for different values of α_t and σ_t^2. Panel B reports the variation in the variance of the nodal strength for different values of γ_t^2 and σ_t^2. For this sensitivity analysis, we assume $N = 100$ and $d_x = 2$.

We can now provide some results characterizing the expected strength and the strength variance of a node i under the assumptions implied by an LS model with an identity link function. The proofs of these propositions can be found in Appendix A.

Proposition 1. *Given Assumptions* 1–4*, the analytical expression of the expected value and variance of the node i strength conditioning on the intercept α_t and the latent coordinate variance σ_t^2 are*

(i) $\mathbb{E}\left(s_{it}|\alpha_t, \sigma_t^2\right) = \sum_{j \neq i} \mathbb{E}\left(w_{ijt}|\alpha_t, \sigma_t^2\right) = (N-1)\left(\alpha_t - 2d_x\sigma_t^2\right)$;

(ii) $\mathbb{V}\left(s_{it}|\alpha_t, \sigma_t^2\right) = (N-1)(\gamma_t^2 + 8\sigma_t^4 d_x) + 2(N-1)(N-2)\sigma_t^4 d_x$.

The expected strength linearly increases as the intercept α_t increases, while it decreases as the variance of the latent coordinates and the dimensions of the latent space increase. The variance of the nodal strength linearly increases with γ_t^2, while it increases exponentially as the variance of the latent coordinates σ_t^2 increases.

From Proposition 1, it is possible to derive the first two moments of the average strength, which can be used to obtain a normal approximation of the weighted degree distribution.

Proposition 2. *The conditional mean and variance of the average strength $\bar{s}_t$ are*

(i) $\mathbb{E}\left(\bar{s}_t|\alpha_t, \sigma_t^2\right) = \alpha_t(N-1) - 2d_x(N-1)\sigma_t^2$;

(ii) $\mathbb{V}\left(\bar{s}_t|\alpha_t, \sigma_t^2\right) = \frac{2(N-1)}{N}\left(\gamma_t^2 + 4N\sigma_t^4 d_x\right)$.

Figure 2 shows how the expected strength and strength's variance change as the latent coordinates' variance σ_t^2 increases for different values of α_t and γ_t^2. As the prior variance of the latent coordinates gets larger, the nodes get further apart on the latent space, and the connections across nodes get more negative. Also, heterogeneity in nodal strength increases with larger variance. Moreover, a larger intercept α_t corresponds to a larger expected strength while a larger γ_t^2 is associated with a larger strength's variability. This effect can hardly be detected for large values of σ_t^2.

3 Bayesian Estimation

The complexity of the LS model naturally calls for a Bayesian estimation procedure. Sampling the latent coordinates from an infinite mixture of normal distributions requires a specific prior choice for the parameters

$\boldsymbol{\phi}_t = \{\boldsymbol{\mu}_t, \sigma_t^2\}$. Following a non-parametric approach, we choose $\boldsymbol{\phi}_t | P_t \sim P_t(\boldsymbol{\phi}_t)$, where P_t for $t = 1, \dots, T$ is a sequence of random probability measures drawn from a Dirichlet Process with concentration parameter ψ_t and base measure P_0 – that is $P_t | \psi_t \sim DP(\psi_t, P_0)$. Such a Dirichlet Process prior admits the following representation:

$$P_t(\mathrm{d}\boldsymbol{\phi}_t) = \sum_{k=1}^{\infty} \lambda_{t,k} \delta_{\boldsymbol{\phi}_{t,k}}(\mathrm{d}\boldsymbol{\phi}_t), \tag{6}$$

where $\lambda_{t,k}$ are the random weights generated by the stick-breaking construction $\lambda_{t,k} = \eta_{t,k} \prod_{l=1}^{k-1}(1 - \eta_{t,l})$ in which each $\eta_{t,l}$ is randomly drawn from a beta distribution with parameters 1 and ψ_t – that is $\eta_{t,l} \sim \mathcal{B}e(1, \psi_t)$ – while the base measure P_0 is given by the product of $\mathcal{N}_{d_x}\left(\mathbf{0}, \omega^2 I_{d_x}\right)$ and $\mathcal{IG}(a_\sigma, b_\sigma)$. We endow the log-concentration $\log \psi_t = h_t$ with the following random-walk dynamics:

$$h_t = h_{t-1} + \xi_t, \quad \xi_t \sim \mathcal{N}\left(0, \sigma_\xi^2\right), \tag{7}$$

$t = 1, \dots, T$, where $h_0 \sim \mathcal{N}\left(0, \sigma_{h_0}^2\right)$.

Finally, we propose an inverse-gamma prior for the variance parameter γ_t^2, that is $\gamma_t^2 \sim \mathcal{IG}(a_\gamma, b_\gamma)$. We summarize the model through the Directed Acyclic Graph reported in Figure 3.

3.1 Posterior Approximation

Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \boldsymbol{\vartheta})$ be the set of parameters. The likelihood of our model can be written as:

$$f(\mathbf{w}|\boldsymbol{\Theta}) = \prod_{t=1}^{T} \prod_{i<j} \mathcal{K}\left(w_{ijt} | \nu_{ijt}, \gamma_t^2\right),$$

where $\boldsymbol{\theta} = \left(\alpha_t, \gamma_t^2, \mathbf{x}_t, \boldsymbol{\mu}_{t,k}, \sigma_{t,k}^2, \psi_t\right)_{t=1}^{T}$ denotes the collection of time-varying parameters, $\boldsymbol{\vartheta} = \left(\sigma_{\alpha_0}^2, \sigma_\epsilon^2\right)$ denotes the set of time-invariant parameters and $\mathcal{K}\left(w_{ijt} | \nu_{ijt}, \gamma_t^2\right)$ denotes the density function of a normal distribution

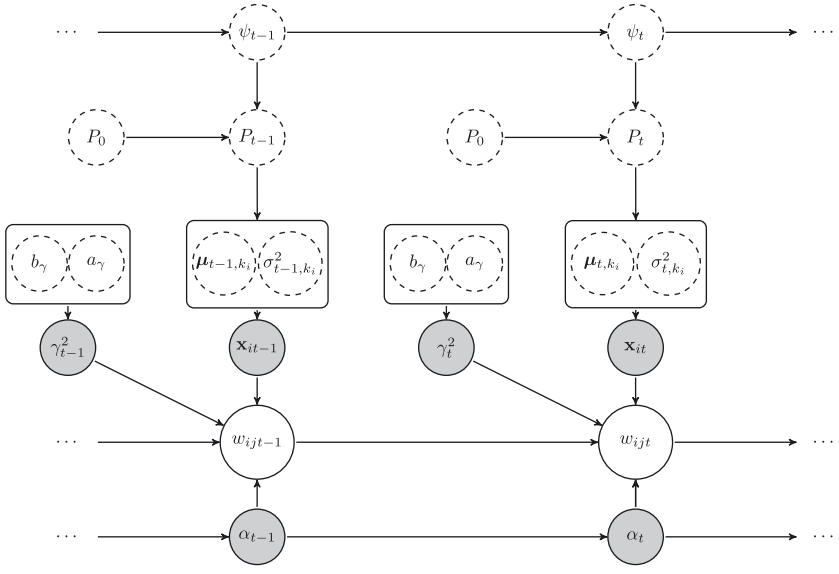

Figure 3: Directed acyclic graph of the proposed dynamic LS model. It exhibits the hierarchical structure of the observations w_{ijt} (solid white circle), the intercept α_t, the latent positions $\mathbf{x}_{it}$ and variances γ_t^2 (gray circles), the parameters of the LS model $\boldsymbol{\mu}_{t,k_i}$ and σ_{t,k_i}^2, the Dirichlet process prior at time t, $DP(\psi_t, P_0)$, denoted with P_t and its hyper-parameters (dashed white circle). The directed arrows show the causal dependence structure of the model.

with mean ν_{ijt} and variance γ_t^2. The joint posterior distribution stemming from the likelihood combined with the priors described above is not analytically tractable. Thus we approximate the joint posterior distribution by Markov-chain Monte Carlo (MCMC). Sampling for the Bayesian nonparametric component of the model is carried out by using the slice-sampling procedure proposed in Walker (2007) and Kalli, Griffin, and Walker (2011).

Denote with K_t the total number of clusters, $N_{t,k}$ the number of nodes (assets) assigned to cluster k, $\bar{\mathbf{x}}_{t,k}$ the vector of centroids of the latent coordinates in cluster k, and $v_{t,k}^2 = \left(\sum_{i=1}^{N_{t,k}} (\mathbf{x}_{it} - \boldsymbol{\mu}_{t,k})^T (\mathbf{x}_{it} - \boldsymbol{\mu}_{t,k})\right)/2$. Then, for each iteration, our MCMC algorithm accomplishes the following steps.

(a) For each time interval t from 1 to T:

1 Draw $\mathbf{x}_{it}$ from $\pi(\mathbf{x}_{it}| \dots)$ via Random-Walk Metropolis Hastings (RW-MH).
2 Draw α_t from $\pi(\alpha_t| \dots)$ via RW-MH.
3 Draw γ_t^2 from $\pi\left(\gamma_t^2| \dots\right)$ via RW-MH.
4 Draw $h_t = \log \psi_t$ from $\pi(h_t| \dots)$ via RW-MH.
5 Draw $\lambda_{t,k}$ and K_t by means of the stick breaking procedure in Walker (2007).
6 Draw exactly

$$(\boldsymbol{\mu}_{t,k}| \dots) \sim \mathcal{N}_{d_x}\left(\frac{N_{t,k}\bar{\mathbf{x}}_{t,k}}{N_{t,k} + \sigma_{t,k}^2/\omega^2}, \frac{\sigma_{t,k}^2}{N_{t,k} + \sigma_{t,k}^2/\omega^2} I_{d_x}\right),$$

with $k \in \{1, \dots, K_t | N_{t,k} \neq 0\}$, otherwise sample from the prior.

7 Draw exactly

$$\left(\sigma_{t,k}^2| \dots\right) \sim \mathcal{IG}\left(a_\sigma + \frac{N_{t,k} d_x}{2}, b_\sigma + v_{t,k}^2\right),$$

with $k \in \{1, \dots, K_t | N_{t,k} \neq 0\}$, otherwise, sample from the prior.

8 Allocate latent coordinates to clusters as in Walker (2007).

(c) Draw exactly $\left(\sigma_\epsilon^2| \dots\right) \sim \mathcal{IG}\left(a_\epsilon + \frac{1}{2}(T-1), b_\epsilon + \frac{1}{2}\sum_{t=2}^{T}\left(\alpha_t - \alpha_{t-1}\right)^2\right)$.

(d) Draw exactly $\left(\sigma_{\alpha_0}^2| \dots\right) \sim \mathcal{IG}\left(a_{\alpha_0} + \frac{1}{2}, b_{\alpha_0} + \frac{1}{2}\alpha_1^2\right)$.

3.2 Likelihood Invariance and Post-Processing

The MCMC output we obtain suffers from two main issues. The first issue is due to the fact that latent coordinates enter the parameter ν_{ijt} only through the squared distance. Thus, positions that differ just by means of reflection, translation, and rotation are equally likely. To overcome this identification issue, it is common to apply a Procrustes transformation to the sampled positions with respect to a set of reference positions. A full description of such a procedure can be found in Hoff, Raftery, and Handcock (2002) and Friel et al. (2016). For example, one may consider as reference positions those returning the highest value for the likelihood.

The second issue is a two-fold label-switching problem. On the one hand, the same cluster may fall under a different label in different MCMC iterations. Moreover, the number of clusters may vary from one iteration to another as a side-effect of the flexibility of our model. We address this issue by applying the Equivalence Classes Representatives algorithm to the sampled allocations as suggested by Papastamoulis (2014).

3.3 Numerical Efficiency Study

We assess the efficiency and effectiveness of the proposed MCMC algorithm through simulation experiments. To do so, we assume to observe the latent position of 15 assets in five epochs. The latent positions of the assets are drawn from a normal distribution centered at (0, 0) in epoch 1 and epoch 3 (single cluster), while the assets are split into groups of equal size and their latent positions are respectively centered at $\{(-1,-1),(0,1),(1,-1)\}$ – thus forming three clusters on the plane – in epochs 2, 4 and 5. The intercept α_t for $t = 1, \dots, 5$ is assumed to

follow a random walk process starting at $\alpha_0 = 5$ with increments $\epsilon_t \overset{iid}{\sim} \mathcal{N}(0,1)$. With the aforementioned parameter specification and the LS mode described in Section 2, we can generate a synthetic dataset on which we can test our algorithm. Figure 4 summarizes the setup of the simulation exercise. Panel A reports the true and the estimated intercept parameter α_t. Panel B reports the position of our 15 assets on the latent plane in the five epochs.

Figure 5 reports the 95 % credible region of the time-varying intercept α_t (shaded-area) against the ground truth (black line). The intercept is correctly estimated. Figure 6 summarises the results at epoch 1. Comparing the true and the estimated values, one can see that the inference procedure provides satisfactory estimation results for the latent coordinates (Panel A), the number of clusters (Panel B), as well as for the other model parameters (Panel C). Trace plots depicting the converge results of our algorithm are reported in the Appendix (see Figure 10).

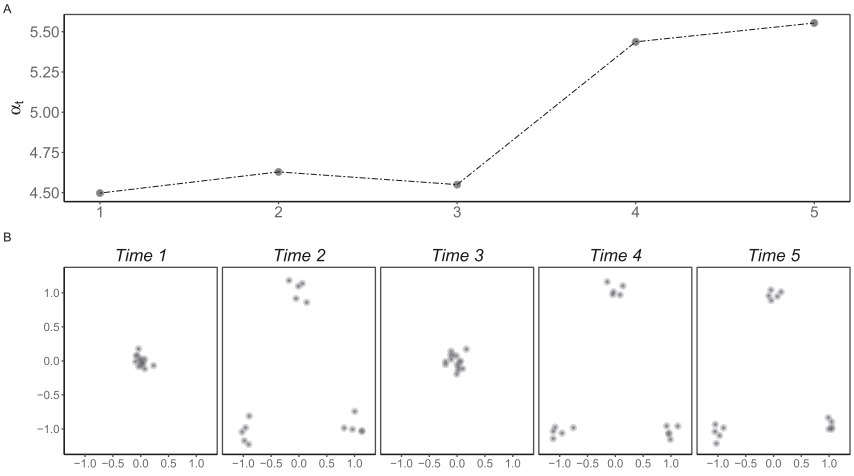

Figure 4: Synthetic data. Time series of the true values of the intercept parameter α_t (top) and of the positions of our 15 synthetic assets over 5 periods (different plots).

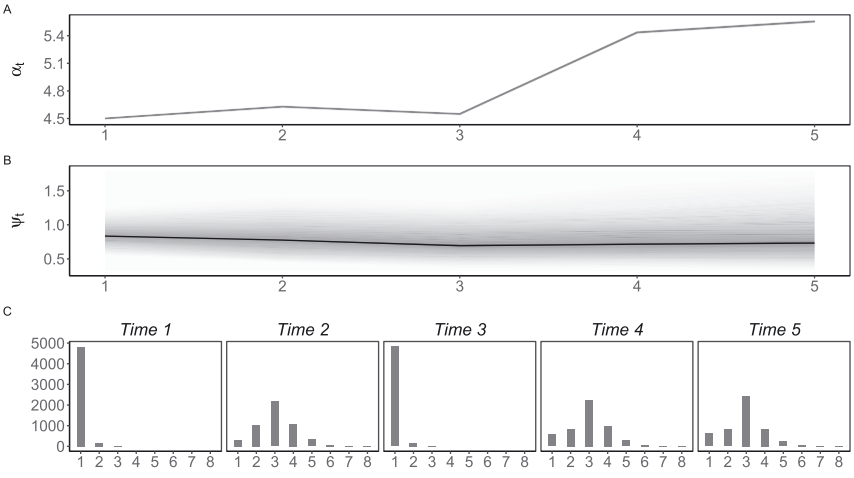

Figure 5: Model simulation results. Sequence of the 95 % credible intervals (shaded area) of the intercept parameter α_t (top) and concentration parameter ψ_t (middle) over time (horizontal axis). Posterior distribution of the number of clusters (bottom) for the five periods in simulation. Note that the posterior mode matches with the correct number of clusters in Figure 4.

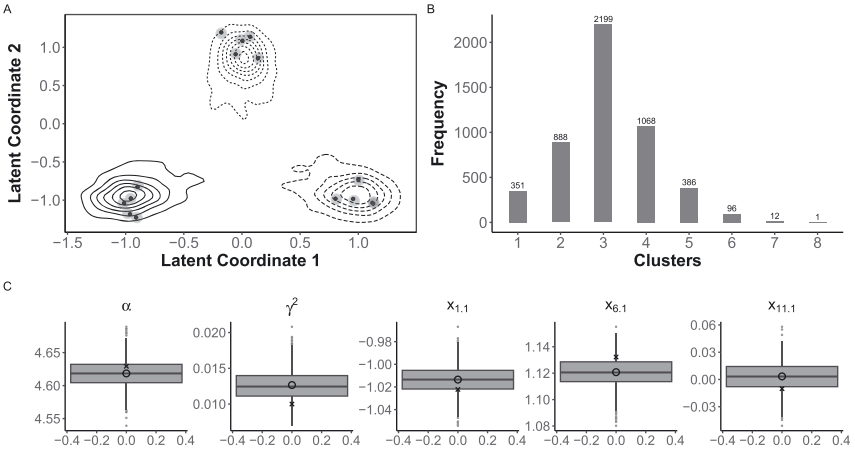

Figure 6: Model simulation results – epoch 2 with a starting setup of 10,000 iterations, after discarding the first 5000 observations as burn-in. Panel A: latent-position estimates with MCMC draws in black dots and reference positions in grey, while contour lines represent the frequency of the draws for $\boldsymbol{\mu}_{2,1}\boldsymbol{\mu}_{2,2}$, and $\boldsymbol{\mu}_{2,3}$. Panel B: bar plot representing the most frequently estimated number of clusters. C: box plot depicting the marginal posterior distributions for the parameters α_2, γ_2^2, $\mathbf{x}_{2,1,1}$, $\mathbf{x}_{2,1,6}$, $\mathbf{x}_{2,1,11}$. Circles represent the sample mean of the marginal posterior, while crosses represent the true value of the parameter.

Table 1: Auto-correlation function (ACF) at lags 1, 10 and 20, the effective sample size (ESS) and the convergence diagnostic (CD) as defined in Geweke (1992) for the MCMC sampling of α_2, γ_2^2, h_2, $\mathbf{x}_{2,1,1}$, $\mathbf{x}_{2,1,6}$, $\mathbf{x}_{2,1,11}$ respectively on the raw MCMC series after burn-in and on the series after burn-in and thinning.

Parameter	**Raw series after burn-in in epoch 2 (5000 obs.)**					
	α_2	γ_2^2	h_2	$\mathbf{x}_{2,1,1}$	$\mathbf{x}_{2,1,6}$	$\mathbf{x}_{2,1,11}$
ACF(1)	0.902	0.821	0.91	0.825	0.872	0.8
ACF(10)	0.417	0.178	0.498	0.217	0.324	0.201
ACF(20)	0.186	0.03	0.319	0.083	0.139	0.064
ESS	228	414	176	420	341	459
CD	1.953 (0.025)	0.206 (0.419)	−3.079 (0.001)	1.457 (0.073)	1.186 (0.118)	1.066 (0.143)
Parameter	**Raw series after burn-in and 20-period thinning (251 obs.)**					
	α_2	γ_2^2	h_2	$\mathbf{x}_{2,1,1}$	$\mathbf{x}_{2,1,6}$	$\mathbf{x}_{2,1,11}$
ACF(1)	0.267	0.083	0.366	0.044	0.065	0.115
ACF(10)	0.014	−0.09	0.073	−0.025	−0.093	−0.133
ACF(20)	−0.009	0.085	0.02	0.049	0.064	0.009
ESS	118	251	116	251	251	198
CD	1.09 (0.138)	0.634 (0.263)	−1.042 (0.149)	1.901 (0.029)	1.191 (0.117)	1.235 (0.108)

Our MCMC algorithm is run for 10,000 iterations and the first 5000 iterations are discarded as burn-in samples. Table 1 reports the diagnostics concerning the series of MCMC draws at epoch 2. Applying burn-in and thinning helps reduce auto-correlation and increase the effective sample size. On average, the effective sample size and the p-values of the CD statistic, as defined in Geweke (1992), are satisfactory.

4 A Stock Market Application

4.1 Data Description

By means of the identity-link LS infinite-mixture model, we investigate the changes in the clustering structure of the constituents of the DAX40 and the S&P100 as of December 2021. The US and German markets are selected

as their assets are often reference assets for many investors aiming at a global asset allocation. The features of these markets have been analyzed in many papers (e.g. see Chesnay and Jondeau 2001; Conlon, Ruskin, and Crane 2009; Heiberger 2014). Our analysis contributes to this literature by providing a new econometric model for the analysis of the asset connectivity structure.

Daily closing prices for the DAX40 and S&P100 components are obtained for the period ranging from January 2007 to December 2021. We drop from the dataset all the components with missing observations and obtain 29 price series for the DAX40 and 88 price series for the S&P100. We selected 2007 as the starting year of our analysis as it allows for considering the 2008–2009 financial crisis while keeping the panel balanced across markets. For each year t from 2007 to 2021, the correlation matrix $\hat{\mathbf{C}}_t$ among log-return series is estimated. Figure 7 reports two of these correlation matrices for the DAX40 (years 2019 and 2020). We can appreciate the change in average cross-correlation in correspondence to the Covid-19 breakout (dark grey kicks in 2020).

4.2 Results for the DAX40

The elements of $\hat{\mathbf{C}}_t$ belong to the interval $[-1, 1]$ and are mapped into the set of the reals, $\mathbb{R}$, by a monotone transformation. We apply here the Z-transformation (see Appendix C.1 for further details) to each entry and fit our LS infinite-mixture model on the sequence of transformed correlations. As an illustrative example, Panel A in Figure 8 reports the latent coordinates $\mathbf{x}_{it}$ estimated in the year 2020 for the DAX40. The model detects the presence of two main clusters in the latent positioning of the index components. The great majority of stocks belong to a single cluster, while QIA and SRT3 belong to a second separated cluster. This result is coherent in light of the Covid-19 outbreak. In fact, the stock price of QIA and SRT3 – two chemical companies – behaved counter-cyclically w.r.t. the market, probably discounting their expertise in molecular testing. Panel B in Figure 8 reports the estimated time series for the intercept α_t while Panel C reports the time series representing the maximum-a-posteriori number of clusters at each point in time. From these two charts, it emerges that as average network connectivity increases (higher intercept), a higher number of clusters emerges. However, as higher correlation network connectivity is often associated with periods of financial turmoil, we thus have that also the clustering structure varies in periods of financial distress. This seems to be true for the years of the financial crisis (2008–2009), the European debt crisis (2011), the stock market selloff (2015), and the Covid-19 crisis (2020).

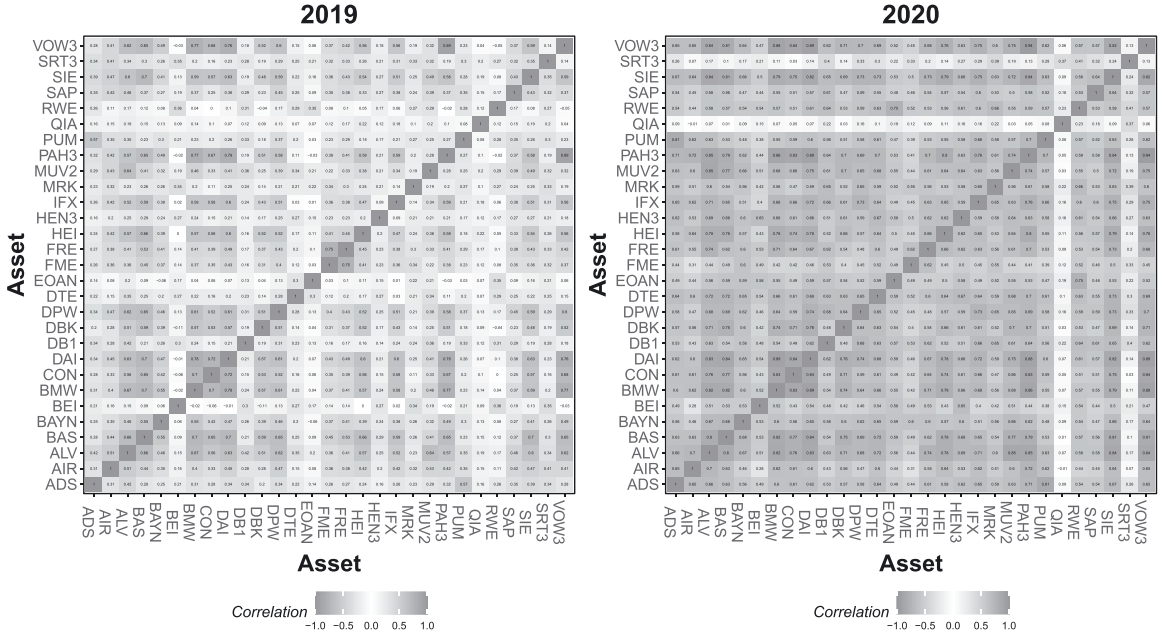

Figure 7: Correlation matrices. Panel A: yearly assets correlation matrix in 2019. Panel B: yearly assets correlation matrix in 2020. Average cross-correlation increased in the year of the Covid-19 breakout.

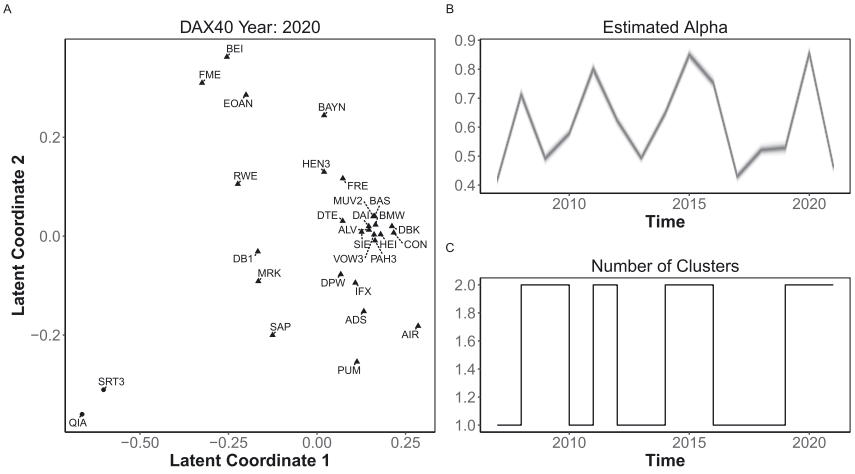

Figure 8: Application – DAX40. Panel A: latent positioning of DAX40 components in the year 2020. The algorithm detects the presence of two clusters (dots and triangles). The majority of the stocks belong to a single cluster, while QIA and SRT3 – two chemicals companies – belong to a second distinct cluster. Panel B: fan chart representing the estimated time series for the intercept α_t and its 95 % credible interval. Panel C: time series representing the MAP estimates of the number of clusters at each point in time.

4.3 Results for the S&P100

Panel A in Figure 9 reports the latent coordinates $\mathbf{x}_{it}$ estimated in the year 2020 for the S&P100. Larger dots correspond to assets with larger market capitalization. The model detects the presence of two main clusters, and there are hints of the presence of a core-periphery structure. Moreover, the fact that larger dots are located at the top of the plane suggests a positive correlation between the second latent coordinate and market capitalization. Panel B in Figure 9 reports the credible intervals for the intercept α_t through time, while Panel C in Figure 9 reports the time series representing the MAP estimates of the number of clusters at each point in time. Also, in

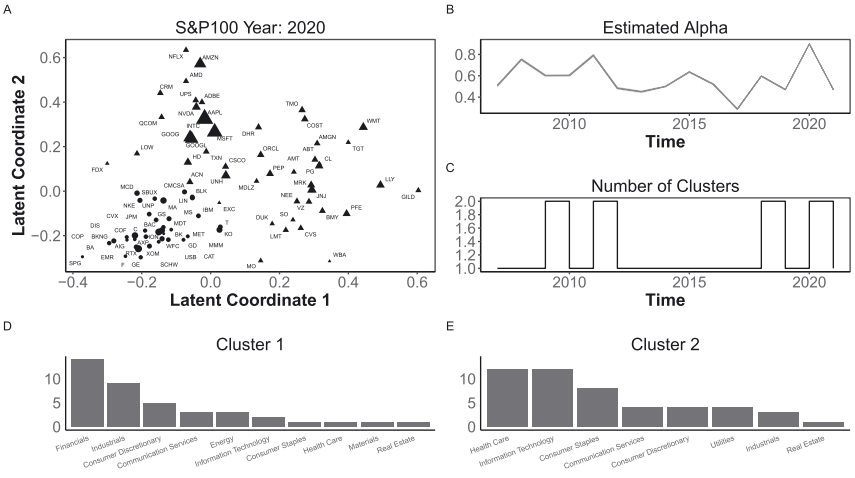

Figure 9: Application – S&P100. Panel A: latent positioning of S & P components in the year 2020. The algorithm detects the presence of two clusters (dots and triangles). Symbols' size is proportional to the market capitalization of the assets. We notice the presence of a core-periphery structure. The core (dots) seems to be composed of financial and industrial assets, while the periphery (triangles) consists of IT, health care, and other sectors. Panel B: fan chart representing the estimated time series for the intercept α_t and its 95 % credible interval. Panel C: time series representing the MAP estimates of the number of clusters at each point in time. Panel D: bar plot reporting the number of assets per sector in cluster 1. Panel E: bar plot reporting the number of assets per sector in cluster 2. Cluster 1, the "core", is mostly populated by assets belonging to the financial and industrial sector, while cluster 2, the "periphery" is mostly populated by IT and healthcare assets.

this case, there are some hints suggesting an association between higher average network connectivity and a change in the clustering structure. We also analyze the composition of nodes belonging to the two clusters. As the bar plots in Panel D and E of Figure 9 suggest, the core (black dots) is mostly populated by Financial and Industrial assets, while the periphery (black triangles) consists of IT, Healthcare and other sectors.

4.4 Number of Clusters and Latent Coordinates

We perform a correlation analysis associating the time-varying number of clusters (Panels C in Figures 8 and 9) with both market volatility and average cross-correlation. The results are reported in Table 2. We notice how the number of clusters correlates positively with both time series, confirming the intuition according to which the assets' clustering structure changes, especially in periods of financial distress. This result remains valid even when considering longer time spans (see Appendix C.2 for more details).

We perform a further correlation analysis on the relationship between the latent coordinates and two relevant financial factors, namely market capitalization and volatility. Table 3 reports the Pearson correlation between the latent coordinates and both Market Capitalization and Volatility in the year 2020 for the DAX40 and the S&P100. In the case of volatility, we notice how at least one of the two coordinates exhibits a sufficiently strong negative correlation with such a factor. Results are mixed for what concerns the correlation between latent coordinates and market capitalization. DAX40's latent coordinates show a weak negative correlation with market capitalization, while the opposite is true for the S&P100. Such correlation provides hints on the existence of some relationship between the latent factors and assets' features.

To summarise, we provided a new model to project the correlation between assets onto a latent space and detect the asset clustering structure. The latent projection conveys information about cross-asset interdependence. The further apart a stock is located on the latent space, the lower its positive correlation with the others. Clusters (groups of assets) identify stocks with similar interdependence features. We applied the novel LS framework to two major reference markets for global investors and a time frame that includes several periods of financial distress.

Cross-sectional clustering and time-varying clustering have been previously addressed in the literature. Fisher and Jensen (2022b) apply the Bayesian Non-parametric (BNP) factor model to a panel of equity funds and find the population has two clusters with different skills. Casarin, Costantini, and Osuntuyi (2023) apply a BNP GARCH model to the S&P100 constituents and find clear-cut evidence of time-varying asset clustering, with clusters having heterogeneous composition across sectors and style features. These results are in line with our findings. Nonetheless, our dynamic LS model allows not only for the detection of clusters but also for a

Table 2: Correlation between the estimated number of clusters and both annualized volatility and average cross-correlation over time for DAX40 and S&P100.

	Volatility		Average cross-correlation	
	DAX40	**S&P100**	**DAX40**	**S&P100**
Median number of clusters	+54.4 %	+48 %	+24.5 %	+51 %

Table 3: Correlation between the latent coordinates and both market capitalization and volatility in the year 2020 for the DAX40 and the S&P100.

	DAX40		S&P100	
	MarketCap	**Volatility**	**MarketCap**	**Volatility**
Latet coordinate 1	−15.3 %	−14.9 %	+0.39 %	−68.6 %
Latent coordinate 2	−10,5 %	−47 %	+36.1 %	−34 %

fine-grained analysis of assets' interdependence, taking advantage of the planar representation of the assets. Moreover, we were able to connect the cross-sectional clustering to a different dimension of risk, which is time-varying interdependence measured as the average pairwise cross-correlation across the set of assets. In periods of financial turmoils, the number of clusters increases and goes in pairs with average cross-correlation. This is in line with what other works suggested (Ahelegbey, Carvalho, and Kolaczyk 2020; Christodoulakis 2007), although, here, the focus is on the effective number of clusters rather than on a synthetic metric for clustering. We also provide some hints concerning the relationship between latent coordinates, market capitalization, and volatility. We hypothesize that this result has some connection with those studies analyzing the PCA decomposition of Assets Correlation Matrices (e.g. see Kim and Jeong 2005).

Our results are consistent across markets and are robust to sample-period selection. Our findings shed light on the market's behavior and have relevant empirical implications. The study of similarities in large panels of financial assets is central for investors aiming at risk diversification and portfolio protection and for policy-makers monitoring financial stability and systemic risk. Further research should try to address more in-depth the financial interpretation of latent coordinates and provide tools to effectively exploit such latent factors for portfolio allocation purposes.

5 Conclusions

In this work, we proposed a new dynamic Latent-Space infinite-mixture model. The latent positions represent financial assets on a latent space, using the information embedding in return cross-correlation. The mixture model assumption accounts for clustering effects in the cross-section of assets, and the infinite components prior assumption allows for estimating the number of clusters. We believe that our model can be considered a novel and relevant tool for analyzing similarities across financial assets. An analysis of such sort seems appealing to investors and policymakers. While the former may exploit our model for portfolio diversification purposes, the latter may consider it an auxiliary tool for monitoring financial stability. We provided an illustrative application of our model to the constituents of DAX40 and S&P100 and found evidence of time-varying asset clustering. In particular, we provided further evidence corroborating the hypothesis according to which asset clustering seems to increase in periods of financial distress. Such evidence points toward a relationship between contagion effects and market risk. We also provide hints on a possible interpretation of the latent coordinates in terms of market capitalization and asset volatility. Nonetheless, we reckon that further research should tackle the financial interpretation of latent coordinates. On the methodological side, we believe that the seamless integration of LS models with portfolio allocation tools may be considered a promising development.

Acknowledgment: The authors acknowledge support from: the MUR – PRIN project '*Discrete random structures for Bayesian learning and prediction*' under g.a. n. 2022CLTYP4 and the Next Generation EU – '*GRINS – Growing Resilient, INclusive and Sustainable*' project (PE0000018), National Recovery and Resilience Plan (NRRP) – PE9 – Mission 4, C2, Intervention 1.3. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.
Code and datasets: The authors published code and data associated with this article on Code Ocean, a computational reproducibility platform. We recommend Code Ocean to SNDE contributors who wish to share, discover, and run code in published research articles (see: https://doi.org/10.24433/CO.1842159.v1). Code and data are also available on GitHub (see: https://github.com/antonioperuzzi94/SNDE_LSM_AssetClustering).
Research funding: This work was supported by MUR – PRIN project "Discrete random structures for Bayesian learning and prediction" (2022CLTYP4) and Next Generation EU Project GRINS – Growing Resilient, INclusive and Sustainable.

Appendix A: Proofs of the Model Properties

A.1 Preliminary Results

In the following, we introduce some results, which are used in the proof of the main results of the paper. In particular, we provide some useful properties of the latent position variables $\mathbf{x}_{it}$.

Proposition 3. *The fourth-order mixed moments of the variables $\boldsymbol{x}_{it}$ and $\boldsymbol{x}_{jt}$, $i \neq j$ are*

(a) $\mathbb{E}\left(\mathbf{x}'_{it}\mathbf{x}_{it}\mathbf{x}'_{it}\mathbf{x}_{it}|\alpha_t,\sigma_t^2\right) = \mathbb{E}\left(\mathbf{x}'_{jt}\mathbf{x}_{jt}\mathbf{x}'_{jt}\mathbf{x}_{jt}|\alpha_t,\sigma_t^2\right) = 3\sigma_t^4 d_x + \sigma_t^4 d_x(d_x-1)$

(b) $\mathbb{E}\left(\mathbf{x}'_{it}\mathbf{x}_{it}\mathbf{x}'_{it}\mathbf{x}_{jt}|\alpha_t,\sigma_t^2\right) = \mathbb{E}\left(\mathbf{x}'_{jt}\mathbf{x}_{jt}\mathbf{x}'_{jt}\mathbf{x}_{it}|\alpha_t,\sigma_t^2\right) = 0$

(c) $\mathbb{E}\left(\mathbf{x}'_{it}\mathbf{x}_{jt}\mathbf{x}'_{it}\mathbf{x}_{it}|\alpha_t,\sigma_t^2\right) = \mathbb{E}\left(\mathbf{x}'_{jt}\mathbf{x}_{it}\mathbf{x}'_{jt}\mathbf{x}_{jt}|\alpha_t,\sigma_t^2\right) = 0$

(d) $\mathbb{E}\left(\mathbf{x}'_{it}\mathbf{x}_{it}\mathbf{x}'_{jt}\mathbf{x}_{jt}|\alpha_t,\sigma_t^2\right) = \mathbb{E}\left(\mathbf{x}'_{jt}\mathbf{x}_{jt}\mathbf{x}'_{it}\mathbf{x}_{it}|\alpha_t,\sigma_t^2\right) = d_x^2\sigma_t^4$

(e) $\mathbb{E}\left(\mathbf{x}'_{it}\mathbf{x}_{jt}\mathbf{x}'_{it}\mathbf{x}_{jt}|\alpha_t,\sigma_t^2\right) = \mathbb{E}\left(\mathbf{x}'_{jt}\mathbf{x}_{it}\mathbf{x}'_{jt}\mathbf{x}_{it}|\alpha_t,\sigma_t^2\right) = d_x\sigma_t^4.$

The fourth-order mixed moments of the variables $\boldsymbol{x}_{it}$, $\boldsymbol{x}_{jt}$, $\boldsymbol{x}_{i't}$, and $\boldsymbol{x}_{j't}$ with distinct subscript indexes i, i', j, j' are:

(f) $\mathbb{E}\left(\mathbf{x}'_{it}\mathbf{x}_{it}\mathbf{x}'_{i't}\mathbf{x}_{i't}|\alpha_t,\sigma_t^2\right) = \mathbb{E}\left(\mathbf{x}'_{it}\mathbf{x}_{it}\mathbf{x}'_{j't}\mathbf{x}_{j't}|\alpha_t,\sigma_t^2\right) = d_x^2\sigma_t^4$

(g) $\mathbb{E}\left(\mathbf{x}'_{jt}\mathbf{x}_{jt}\mathbf{x}'_{i't}\mathbf{x}_{i't}|\alpha_t,\sigma_t^2\right) = \mathbb{E}\left(\mathbf{x}'_{jt}\mathbf{x}_{jt}\mathbf{x}'_{j't}\mathbf{x}_{j't}|\alpha_t,\sigma_t^2\right) = d_x^2\sigma_t^4$

(h) $\mathbb{E}\left(\mathbf{x}'_{it}\mathbf{x}_{it}\mathbf{x}'_{i't}\mathbf{x}_{j't}|\alpha_t,\sigma_t^2\right) = \mathbb{E}\left(\mathbf{x}'_{jt}\mathbf{x}_{jt}\mathbf{x}'_{i't}\mathbf{x}_{j't}|\alpha_t,\sigma_t^2\right) = 0$

(i) $\mathbb{E}\left(\mathbf{x}'_{it}\mathbf{x}_{jt}\mathbf{x}'_{i't}\mathbf{x}_{i't}|\alpha_t,\sigma_t^2\right) = \mathbb{E}\left(\mathbf{x}'_{it}\mathbf{x}_{jt}\mathbf{x}'_{j't}\mathbf{x}_{j't}|\alpha_t,\sigma_t^2\right) = 0$

(j) $\mathbb{E}\left(\mathbf{x}'_{it}\mathbf{x}_{jt}\mathbf{x}'_{i't}\mathbf{x}_{jt}|\alpha_t,\sigma_t^2\right) = 0.$

Proof. Follows from the independee nce and normal assumptions (Assumptions 1 and 3) and the properties of the multivariate normal distribution.

Some useful properties of the strength variables w_{ijt} are given in the following.

Proposition 4. *The first and second order conditional moments of the strength variables w_{ijt} given α_t and σ_t^2 are:*

(i) $\mathbb{E}\left(w_{ijt}|\alpha_t,\sigma_t^2\right) = \alpha_t - 2d_x\sigma_t^2$

(ii) $\mathbb{V}\left(w_{ijt}|\alpha_t,\sigma_t^2\right) = \gamma_t^2 + 8\sigma_t^4 d_x$

(iii) $\mathbb{C}ov\left(w_{ijt}, w_{i'j't}|\alpha_t,\sigma_t^2\right) = 0$

(iv) $\mathbb{C}ov\left(w_{ijt}, w_{i'jt}|\alpha_t,\sigma_t^2\right) = 2\sigma_t^4 d_x$

Proof.

(i) By the law of iterated expectations and Assumptions 1 and 3 one gets:

$$\begin{aligned}
\mathbb{E}\left(w_{ijt}|\alpha_t,\sigma_t^2\right) &= \mathbb{E}\left(\mathbb{E}\left(w_{ijt}|\alpha_t,\sigma_t^2,\mathbf{x}_t\right)|\alpha_t,\sigma_t^2\right) \\
&= \alpha_t - \mathbb{E}\left(||\mathbf{x}_{it}-\mathbf{x}_{jt}||^2|\alpha_t,\sigma_t^2\right) \\
&= \alpha_t - \int_{\mathbf{x}_{it}}\int_{\mathbf{x}_{jt}} (\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt}) f(\mathbf{x}_{it}) f(\mathbf{x}_{jt}) \mathrm{d}\mathbf{x}_{it}\mathrm{d}\mathbf{x}_{jt} \\
&= \alpha_t - \left(\int_{\mathbf{x}_{it}} \mathbf{x}'_{it}\mathbf{x}_{it} f(\mathbf{x}_{it})\mathrm{d}\mathbf{x}_{it} + \int_{\mathbf{x}_{jt}} \mathbf{x}'_{jt}\mathbf{x}_{jt} f(\mathbf{x}_{jt})\mathrm{d}\mathbf{x}_{it}\right)
\end{aligned}$$

$$= \alpha_t - 2d_x\sigma_t^2$$

(ii) We can write the variance of the weight given α_t and σ_t^2 as:

$$\begin{aligned}
\mathbb{V}(w_{ijt}|\alpha_t,\sigma_t^2) &= \mathbb{E}\big(\mathbb{V}(w_{ijt}|\alpha_t,\sigma_t^2,\mathbf{x}_t)|\alpha_t,\sigma_t^2\big) + \mathbb{V}(\mathbb{E}(w_{ijt}|\alpha_t,\sigma_t^2,\mathbf{x}_t)|\alpha_t,\sigma_t^2)\\
&= \gamma_t^2 + \mathbb{V}\big(\alpha_t - ||\mathbf{x}_{it}-\mathbf{x}_{jt}||^2|\alpha_t,\sigma_t^2\big)\\
&= \gamma_t^2 + \mathbb{V}((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt})|\alpha_t,\sigma_t^2)\\
&= \gamma_t^2 + \underbrace{\mathbb{E}(((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt}))^2|\alpha_t,\sigma_t^2)}_{=A} - \underbrace{\big[\mathbb{E}((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt}))|\alpha_t,\sigma_t^2)\big]^2}_{=B}
\end{aligned}$$

Simple algebra allows us to decompose A as follows:

$$\begin{aligned}
A &= \mathbb{E}(((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt}))^2|\alpha_t,\sigma_t^2)\\
&= \mathbb{E}\Big(\mathbf{x}'_{it}\mathbf{x}_{it}\mathbf{x}'_{it}\mathbf{x}_{it} + \mathbf{x}'_{it}\mathbf{x}_{it}\mathbf{x}'_{jt}\mathbf{x}_{jt} - 2\mathbf{x}'_{it}\mathbf{x}_{it}\mathbf{x}'_{it}\mathbf{x}_{jt}\\
&\quad + \mathbf{x}'_{jt}\mathbf{x}_{jt}\mathbf{x}'_{it}\mathbf{x}_{it} + \mathbf{x}'_{jt}\mathbf{x}_{jt}\mathbf{x}'_{jt}\mathbf{x}_{jt} - 2\mathbf{x}'_{jt}\mathbf{x}_{jt}\mathbf{x}'_{it}\mathbf{x}_{jt}\\
&\quad - 2\mathbf{x}'_{it}\mathbf{x}_{jt}\mathbf{x}'_{it}\mathbf{x}_{it} - 2\mathbf{x}'_{it}\mathbf{x}_{jt}\mathbf{x}'_{jt}\mathbf{x}_{jt} + 4\mathbf{x}'_{it}\mathbf{x}_{jt}\mathbf{x}'_{it}\mathbf{x}_{jt}|\alpha_t,\sigma_t^2\Big)
\end{aligned}$$

Following the results in Proposition 3 and after some computations one obtains: $A = 2(3\sigma_t^4 d_x + \sigma_t^4 d_x(d_x-1)) + 2\big(\sigma_t^4 d_x^2\big) + 4\big(\sigma_t^4 d_x\big) = 6\sigma_t^4 d_x + 2\sigma_t^4 d_x^2 - 2\sigma_t^4 d_x + 2\sigma_t^4 d_x^2 + 4\sigma_t^4 d_x = 8\sigma_t^4 d_x + 4\sigma_t^4 d_x^2$. We proceed by solving $\sqrt{B}$:

$$\begin{aligned}
\sqrt{B} &= \mathbb{E}((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt}))|\alpha_t,\sigma_t^2)\\
&= \mathbb{E}(\Big(\mathbf{x}'_{it}\mathbf{x}_{it} + \mathbf{x}'_{jt}\mathbf{x}_{jt} - 2\mathbf{x}'_{it}\mathbf{x}_{jt}\Big)|\alpha_t,\sigma_t^2)\\
&= 2d_x\sigma_t^2.
\end{aligned}$$

We thus find a closed-form solution for the variance: $\mathbb{V}(w_{ijt}|\alpha_t,\sigma_t^2) = \gamma_t^2 + A - B = \gamma_t^2 + 8\sigma_t^4 d_x + 4\sigma_t^4 d_x^2 - 4\sigma_t^4 d_x^2 = \gamma_t^2 + 8\sigma_t^4 d_x$.

(iii) By the law of total covariance:

$$\begin{aligned}
\mathbb{C}ov\big(w_{ijt}, w_{i'j't}|\alpha_t,\sigma_t^2\big) &= \mathbb{E}\big(\mathbb{C}ov\big(w_{ijt}, w_{i'j't}|\alpha_t,\sigma_t^2,\mathbf{x}_t\big)|\alpha_t,\sigma_t^2\big)\\
&\quad + \mathbb{C}ov(\mathbb{E}\big(w_{ijt}|\alpha_t,\sigma_t^2,\mathbf{x}_t\big), \mathbb{E}\big(w_{i'j't}|\alpha_t,\sigma_t^2,\mathbf{x}_t\big)|\alpha_t,\sigma_t^2)\\
&= \mathbb{E}\big(0|\alpha_t,\sigma_t^2\big) + \mathbb{C}ov(\mathbb{E}\big(w_{ijt}|\alpha_t,\sigma_t^2,\mathbf{x}_t\big), \mathbb{E}\big(w_{i'j't}|\alpha_t,\sigma_t^2,\mathbf{x}_t\big)|\alpha_t,\sigma_t^2)\\
&= \mathbb{C}ov((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt}), \big(\mathbf{x}_{i't}-\mathbf{x}_{j't}\big)'\big(\mathbf{x}_{i't}-\mathbf{x}_{j't}\big)|\alpha_t,\sigma_t^2)\\
&= \underbrace{\mathbb{E}((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt})(\mathbf{x}_{i't}-\mathbf{x}_{j't})'\big(\mathbf{x}_{i't}-\mathbf{x}_{j't}\big)|\alpha_t,\sigma_t^2)}_{=C}\\
&\quad - \underbrace{\big[\mathbb{E}((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt}))|\alpha_t,\sigma_t^2)\big]\big[\mathbb{E}(\big(\mathbf{x}_{i't}-\mathbf{x}_{j't}\big)'\big(\mathbf{x}_{i't}-\mathbf{x}_{j't}\big))|\alpha_t,\sigma_t^2)\big]}_{=D}
\end{aligned}$$

We proceed by solving C. By expanding the quadratic form:

$$
\begin{aligned}
C &= \mathbb{E}((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt})(\mathbf{x}_{i't}-\mathbf{x}_{j't})'\left(\mathbf{x}_{i't}-\mathbf{x}_{j't}\right)|\alpha_t,\sigma_t^2)\\
&= \mathbb{E}\Big(\mathbf{x}_{it}'\mathbf{x}_{it}\mathbf{x}_{i't}'\mathbf{x}_{i't}+\mathbf{x}_{it}'\mathbf{x}_{it}\mathbf{x}_{j't}'\mathbf{x}_{j't}-2\mathbf{x}_{it}'\mathbf{x}_{it}\mathbf{x}_{i't}'\mathbf{x}_{j't}\\
&\quad+\mathbf{x}_{jt}'\mathbf{x}_{jt}\mathbf{x}_{i't}'\mathbf{x}_{i't}+\mathbf{x}_{jt}'\mathbf{x}_{jt}\mathbf{x}_{j't}'\mathbf{x}_{j't}-2\mathbf{x}_{jt}'\mathbf{x}_{jt}\mathbf{x}_{i't}'\mathbf{x}_{j't}\\
&\quad-2\mathbf{x}_{it}'\mathbf{x}_{jt}\mathbf{x}_{i't}'\mathbf{x}_{i't}-2\mathbf{x}_{it}'\mathbf{x}_{jt}\mathbf{x}_{j't}'\mathbf{x}_{j't}+4\mathbf{x}_{it}'\mathbf{x}_{jt}\mathbf{x}_{i't}'\mathbf{x}_{j't}|\alpha_t,\sigma_t^2\Big)
\end{aligned}
$$

which is a decomposition in five terms. Following the results in Proposition 3, and after some algebra, we obtain: $C = 4d_x^2\sigma_t^4$. We proceed by solving D:

$$
\begin{aligned}
D &= \left[\mathbb{E}((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt}))|\alpha_t,\sigma_t^2)\right]\left[\mathbb{E}(\left(\mathbf{x}_{i't}-\mathbf{x}_{j't}\right)'\left(\mathbf{x}_{i't}-\mathbf{x}_{j't}\right))|\alpha_t,\sigma_t^2)\right]\\
&= 4d_x^2\sigma_t^4
\end{aligned}
$$

and find that: $\mathbb{C}ov\left(w_{ijt},w_{i'j't}|\alpha_t,\sigma_t^2\right)=C-D=0$.

(iv) By the law of total covariance:

$$
\begin{aligned}
\mathbb{C}ov\left(w_{ijt},w_{i'jt}|\alpha_t,\sigma_t^2\right) &= \mathbb{E}\left(\mathbb{C}ov\left(w_{ijt},w_{i'jt}|\alpha_t,\sigma_t^2,\mathbf{x}_t\right)|\alpha_t,\sigma_t^2\right)\\
&\quad+\mathbb{C}ov(\mathbb{E}\left(w_{ijt}|\alpha_t,\sigma_t^2,\mathbf{x}_t\right),\mathbb{E}\left(w_{i'jt}|\alpha_t,\sigma_t^2,\mathbf{x}_t\right)|\alpha_t,\sigma_t^2)\\
&= \mathbb{E}\left(0|\alpha_t,\sigma_t^2\right)+\mathbb{C}ov(\mathbb{E}\left(w_{ijt}|\alpha_t,\sigma_t^2,\mathbf{x}_t\right),\mathbb{E}\left(w_{i'jt}|\alpha_t,\sigma_t^2,\mathbf{x}_t\right)|\alpha_t,\sigma_t^2)\\
&= \mathbb{C}ov((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt}),\left(\mathbf{x}_{i't}-\mathbf{x}_{jt}\right)'\left(\mathbf{x}_{i't}-\mathbf{x}_{jt}\right)|\alpha_t,\sigma_t^2)\\
&= \underbrace{\mathbb{E}((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt})(\mathbf{x}_{i't}-\mathbf{x}_{jt})'\left(\mathbf{x}_{i't}-\mathbf{x}_{jt}\right)|\alpha_t,\sigma_t^2)}_{=E}\\
&\quad-\underbrace{\left[\mathbb{E}((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt}))|\alpha_t,\sigma_t^2)\right]\left[\mathbb{E}(\left(\mathbf{x}_{i't}-\mathbf{x}_{jt}\right)'\left(\mathbf{x}_{i't}-\mathbf{x}_{jt}\right))|\alpha_t,\sigma_t^2)\right]}_{=F}
\end{aligned}
$$

Following the same line as in the previous point, we proceed by decomposing F as follows:

$$
\begin{aligned}
F &= \mathbb{E}((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt})(\mathbf{x}_{i't}-\mathbf{x}_{jt})'\left(\mathbf{x}_{i't}-\mathbf{x}_{jt}\right)|\alpha_t,\sigma_t^2)\\
&= \mathbb{E}\Big(\mathbf{x}_{it}'\mathbf{x}_{it}\mathbf{x}_{i't}'\mathbf{x}_{i't}+\mathbf{x}_{it}'\mathbf{x}_{it}\mathbf{x}_{jt}'\mathbf{x}_{jt}-2\mathbf{x}_{it}'\mathbf{x}_{it}\mathbf{x}_{i't}'\mathbf{x}_{jt}\\
&\quad+\mathbf{x}_{jt}'\mathbf{x}_{jt}\mathbf{x}_{i't}'\mathbf{x}_{i't}+\mathbf{x}_{jt}'\mathbf{x}_{jt}\mathbf{x}_{jt}'\mathbf{x}_{jt}-2\mathbf{x}_{jt}'\mathbf{x}_{jt}\mathbf{x}_{i't}'\mathbf{x}_{jt}\\
&\quad-2\mathbf{x}_{it}'\mathbf{x}_{jt}\mathbf{x}_{i't}'\mathbf{x}_{i't}-2\mathbf{x}_{it}'\mathbf{x}_{jt}\mathbf{x}_{jt}'\mathbf{x}_{jt}+4\mathbf{x}_{it}'\mathbf{x}_{jt}\mathbf{x}_{i't}'\mathbf{x}_{jt}|\alpha_t,\sigma_t^2\Big)
\end{aligned}
$$

From the results in Proposition 3 and after some algebra one obtains $E = 3\sigma_t^4d_x^2+3\sigma_t^4d_x+\sigma_t^4d_x(d_x-1)$ $=4\sigma_t^4d_x^2+2\sigma_t^4d_x$ and $F=\left[\mathbb{E}((\mathbf{x}_{it}-\mathbf{x}_{jt})'(\mathbf{x}_{it}-\mathbf{x}_{jt}))|\alpha_t,\sigma_t^2)\right]\left[\mathbb{E}(\left(\mathbf{x}_{i't}-\mathbf{x}_{jt}\right)'\left(\mathbf{x}_{i't}-\mathbf{x}_{jt}\right))|\alpha_t,\sigma_t^2)\right]=4d_x^2\sigma_t^4$. We thus find that $\mathbb{C}ov\left(w_{ijt},w_{i'jt}|\alpha_t,\sigma_t^2\right)=E-F=4\sigma_t^4d_x^2+2\sigma_t^4d_x-4d_x^2\sigma_t^4=2\sigma_t^4d_x$.

A.2 Proof of Proposition 1

From Definition 1 and the results in Proposition 4:

(i) the conditional expectation of the strength in our LS model is:

$$
\mathbb{E}\left(s_{it}|\alpha_t,\sigma_t^2\right)=\sum_{j\neq i}\mathbb{E}\left(w_{ijt}|\alpha_t,\sigma_t^2\right)=\alpha_t(N-1)-2d_x(N-1)\sigma_t^2;
$$

(ii) the conditional strength variance is

$$\begin{aligned}\mathbb{V}(s_{it}|\alpha_t,\sigma_t^2) &= \mathbb{V}\left(\sum_{j\neq i} w_{ijt}\middle|\alpha_t,\sigma_t^2\right)\\ &= \sum_{j\neq i}\mathbb{V}(w_{ijt}|\alpha_t,\sigma_t^2) + \sum_{j\neq i}\sum_{j'\neq i'}\mathbb{C}ov(w_{ij},w_{i'j'}|\alpha_t,\sigma_t^2)\\ &= (N-1)(\gamma_t^2+8\sigma_t^4 d_x) + (N-1)(N-2)2\sigma_t^4 d_x.\end{aligned}$$

A.3 Proof of Proposition 2

From Definition 2 and the results in Proposition 1, we obtain

(i) The expected value of the average strength estimator is

$$\mathbb{E}(\bar{s}_t|\alpha_t,\sigma_t^2) = \frac{1}{N}\sum_{j\neq i}\sum \mathbb{E}(w_{ijt}|\alpha_t,\sigma_t^2) = \alpha_t(N-1) - 2d_x(N-1)\sigma_t^2$$

(ii) The variance of the average strength estimator is

$$\begin{aligned}\mathbb{V}(\bar{s}_t|\alpha_t,\sigma_t^2) &= \frac{1}{N^2}\mathbb{V}\left(\sum_{j\neq i}\sum w_{ijt}\middle|\alpha_t,\sigma_t^2\right) = \frac{4}{N^2}\mathbb{V}\left(\sum_{j>i}\sum w_{ijt}\middle|\alpha_t,\sigma_t^2\right)\\ &= \frac{4}{N^2}\left[\sum_{j>i}\sum\mathbb{V}(w_{ijt}|\alpha_t,\sigma_t^2) + \sum_{j>i}\sum_{j'>i'}\sum\mathbb{C}ov(w_{ij},w_{i'j'}|\alpha_t,\sigma_t^2)\right]\\ &= \frac{4}{N^2}\left[\frac{N(N-1)}{2}(\gamma_t^2+8\sigma_t^4 d_x) + \frac{N(N-1)}{2}(2N-4)2\sigma_t^4 d_x)\right]\\ &= \frac{2}{N^2}\left[N(N-1)(\gamma_t^2+8\sigma_t^4 d_x) + 2N(N-1)(2N-4)\sigma_t^4 d_x)\right]\\ &= \frac{2N(N-1)}{N^2}\left[(\gamma_t^2+8\sigma_t^4 d_x) + 2(2N-4)\sigma_t^4 d_x)\right]\\ &= \frac{2(N-1)}{N}\left[\gamma_t^2+8\sigma_t^4 d_x + 4N\sigma_t^4 d_x - 8\sigma_t^4 d_x\right]\\ &= \frac{2(N-1)}{N}\left[\gamma_t^2+4N\sigma_t^4 d_x\right].\end{aligned}$$

Appendix B: Computational Details

B.1 Proof of the Full Conditional Distributions

In this appendix, we derive the full conditional distributions of the Gibbs Sampler for the infinite-mixture LS model with the identity link function. The Gibbs sampler relies on the stick-breaking representation, and the data augmentation step includes the slice and allocation variables as proposed in Walker (2007) and Kalli, Griffin, and Walker (2011). We handle the infinity mixture problem by introducing a set of auxiliary variables (U_t, S_t, H_t)

in the full posterior so that $\boldsymbol{\theta}^* = \left(\alpha_t, \gamma_t^2, \mathbf{x}_t, \boldsymbol{\mu}_{t,k}, \sigma_{t,k}^2, \psi_t, U_t, K_t, H_t\right)_{t=1}^T$, $\boldsymbol{\vartheta} = \left(\sigma_{\alpha_0}^2, \sigma_\epsilon^2\right)$ and $\boldsymbol{\Theta}^* = \{\boldsymbol{\theta}^*, \boldsymbol{\vartheta}\}$. Introducing such a set of variables allows for a finite representation of the infinite mixture. Given the prior choice described above, the full posterior can be written as:

$$\pi(\boldsymbol{\Theta}^*|\mathbf{w}) \propto \prod_{t=1}^{T}\prod_{i<j}\prod \mathcal{K}\left(w_{ijt}|\nu_{ijt}, \gamma_t^2\right)\mathbb{I}_{\{u_{i,t}<\lambda_{t,s_{i,t}}\}} f_{d_x}\left(\mathbf{x}_{t,s_{i,t}}|\boldsymbol{\mu}_{t,s_{i,t}}, \sigma_{t,s_{i,t}}^2 I_{d_x}\right)$$
$$\cdot \prod_{k>1}\left(\pi(\boldsymbol{\mu}_{t,k})\pi\left(\sigma_{t,k}^2\right)(1-\eta_{t,k})^{\psi_t-1}\right)\pi(\psi_t)\pi(\alpha_t)\pi\left(\gamma_t^2\right)\pi(h_t)\pi(\boldsymbol{\vartheta}),$$

where $U_t = (u_{1,t}, u_{2,t}, \dots, u_{N,t})$, $H_t = (\eta_{1,t}, \eta_{2,t}, \dots)$, $S_t = \{s_{1,t}, \dots, s_{N,t}\}$ is a set of allocation variables. We further define $\mathcal{S}_{t,k} = \{i = 1, \dots, N | s_{i,t} = k\}$ as the set of indexes of the observations allocated to the kth component of the mixture, $\mathcal{S}_t = \{k|\text{card}\left(\mathcal{S}_{t,k}\right) \neq 0\}$ the set of indexes with non-empty mixture components, $\text{card}(\mathcal{S}_t)$ is the cardinality of $\mathcal{S}_t$ and we think to $S_t^* = \sup(\mathcal{S}_t)$ as the number of used components. Moreover, we define the maximum number of components to draw as $K_t = \max\{i = 1, \dots, N | K_{i,t}\}$, where $K_{i,t}$ is the smallest integer such that $\sum_{h=1}^{K_{i,t}} \lambda_{h,t} > 1 - u_{i,t}$. Finally, we can sample from the joint posterior by splitting $H_t = \left(H_t^*, H_t^{**}\right)$, for which $H_t^* = \left(\eta_{1,t}, \dots, \eta_{S_t^*,t}\right)$ and $H_t^{**} = \left(\eta_{S_t^*+1,t}, \dots, \eta_{K_t^*,t}\right)$ and sampling from $\pi\left(H_t^*|\dots\right)$, $\pi(U_t|\dots)$ and $\pi\left(H_t^{**}|\dots\right)$ in order.

Full conditional distribution of H_t^*

From the data augmentation framework we obtain a beta distribution for the stick-breaking components:

$$\pi(\eta_{t,k}|\dots) \propto (1-\eta_{t,k})^{\psi_t + b_{k,t} - 1}\eta_{t,k}^{a_{k,t}}$$

where $a_{k,t} = a + \sum_{i=1}^N \mathbb{I}(s_{i,t} = k)$ is the number coordinates in cluster k and $b_{k,t} = b + N - \sum_{i=1}^N \mathbb{I}(s_{i,t} < k)$ is the number of coordinates in clusters with label smaller or equal than k, with $k = 1, \dots, S_t^*$.

Full conditional distribution of U_t

The conditional distribution of the slice variables is the uniform distribution:

$$\pi(u_{i,t}|\dots) \propto \frac{1}{\lambda_{s_{i,t}}}\mathbb{I}_{\{u_{i,t}<\lambda_{s_{i,t}}\}}, \quad i = 1, \dots, N$$

Full conditional distribution of H_t^{**}

The conditional distribution of the slice variables is the beta distribution:

$$\pi(\eta_{t,k}|\dots) \propto (1-\eta_{t,k})^{\psi_t - 1}.$$

with $k = S_t^* + 1, \dots, K_t$

Full conditional distribution of S_t

$$\pi(s_{i,t}|\dots) = \mathbb{I}_{\{u_{i,t}<\lambda_{s_{i,t}}\}} f_{d_x}\left(\mathbf{x}_{s_{i,t}}|\boldsymbol{\mu}_{s_{i,t}}, \sigma_{s_{i,t}}^2 I_{d_x}\right).$$

Full conditional distribution of h_t

The conditional distribution of the latent log-concentration process

$$\pi(h_t|\dots) \propto \exp(K_t h_t)\frac{\Gamma\left(\exp(h_t)\right)}{\Gamma\left(\exp(h_t) + N\right)}\left[f\left(h_t|0, \sigma_{h_0}^2\right)\right]^{\mathbb{I}(t=1)}\left[f\left(h_t|h_{t-1}, \sigma_\xi^2\right)\right]^{\mathbb{I}(t>1)}$$
$$\times \left[f\left(h_{t+1}|h_t, \sigma_\xi^2\right)\right]^{\mathbb{I}(t<T)},$$

where N is the number of nodes and K_t is the number of mixture components and $f(\cdot)$ denotes a Gaussian pdf. The distribution is not tractable and samples can be obtained by a Metropolis-Hasting step with a Gaussian proposal.

Full conditional distribution of $\mu_{t,k}$

$$\begin{aligned}\pi(\boldsymbol{\mu}_{t,k}|\dots) &\propto \pi(\boldsymbol{\mu}_{t,k})\prod_{i\in S_k} f_{d_x}\left(\mathbf{x}_{t,s_{i,t}}|\boldsymbol{\mu}_{t,s_{i,t}},\sigma^2_{t,s_{i,t}}I_{d_x}\right)\\ &\propto f_{d_x}\left(\boldsymbol{\mu}_{t,k}|0,\omega^2 I_{d_x}\right)\prod_{i\in S_k} f_{d_x}\left(\mathbf{x}_{t,s_{i,t}}|\boldsymbol{\mu}_{t,s_{i,t}},\sigma^2_{t,s_{i,t}}I_{d_x}\right)\\ &\propto \mathcal{N}_{d_x}\left(\frac{N_{t,k}\bar{\mathbf{x}}_k}{N_{t,k}+\sigma^2_{t,k}/\omega^2},\frac{\sigma^2_{t,k}}{N_{t,k}+\sigma^2_{t,k}/\omega^2}I_{d_x}\right),\end{aligned}$$

for $k=1,\dots,K_t$ as in Handcock, Raftery, and Tantrum (2007), where $N_{t,k}$ the number of latent coordinates assigned to cluster k and $\bar{\mathbf{x}}_{t,k}$ is the vector of average latent coordinates in cluster k.

Full conditional distribution of $\sigma^2_{t,k}$

$$\begin{aligned}\pi\left(\sigma^2_{t,k}|\dots\right) &\propto \pi\left(\sigma^2_{t,k}\right)\prod_{i\in S_k} f_{d_x}\left(\mathbf{x}_{t,s_{i,t}}|\boldsymbol{\mu}_{t,d_{i,t}},\sigma^2_{t,s_{i,t}}I_{d_x}\right)\\ &\propto \mathcal{IG}\left(a_\sigma+\frac{N_{t,k}d_x}{2},b_\sigma+v^2_{t,k}\right),\quad k\in\{1,\dots,K_t|card(S_{t,k})\}\end{aligned}$$

for $k=1,\dots,K_t$ as in Handcock, Raftery, and Tantrum (2007), where $N_{t,k}$ the number of latent coordinates assigned to cluster k, $\bar{\mathbf{x}}_{t,k}$ the vector of average latent coordinates in cluster k, and $v^2_{t,k}=(2)^{-1}\sum_{i=1}^{N_{t,k}}(\mathbf{x}_{i,t}-\boldsymbol{\mu}_{t,k})^T(\mathbf{x}_{i,t}-\boldsymbol{\mu}_{t,k})$.

Full conditional distribution of α_t

The full conditional of α_t is sampled via a random walk Metropolis–Hastings.

$$\pi(\alpha_t|\dots)\propto\left(\prod_{i<j}\prod\mathcal{K}\left(w_{ijt}|\nu_{ijt},\gamma^2_t\right)\right)\left(f\left(\alpha_t|0,\sigma^2_{\alpha_0}\right)\right)^{\mathbb{I}(t=1)}\left(f\left(\alpha_t|\alpha_{t-1},\sigma^2_\epsilon\right)\right)^{\mathbb{I}(t>1)}\left(f\left(\alpha_{t+1}|\alpha_t,\sigma^2_\epsilon\right)\right)^{\mathbb{I}(t<T)}.$$

Full conditional distribution of γ^2_t

The full conditional of γ^2_t is sampled via a random walk Metropolis-Hastings.

$$\pi\left(\gamma^2_t|\dots\right)\propto\left(\prod_{i<j}\prod\mathcal{K}\left(w_{ijt}|\nu_{ijt},\gamma^2_t\right)\right)\pi\left(\gamma^2_t\right).$$

Full conditional distribution of σ^2_ϵ

$$\pi\left(\sigma^2_\epsilon|\dots\right)\propto\mathcal{IG}\left(a_\epsilon+\frac{1}{2}(T-1),b_\epsilon+\frac{1}{2}\sum_{t=2}^{T}\left(\alpha_t-\alpha_{t-1}\right)^2\right).$$

Full conditional distribution of $\sigma^2_{\alpha_0}$

$$\pi\left(\sigma^2_{\alpha_0}|\dots\right)\propto\mathcal{IG}\left(a_{\alpha_0}+\frac{1}{2},b_{\alpha_0}+\frac{1}{2}\alpha_1^2\right).$$

B.2 Further Simulation Results

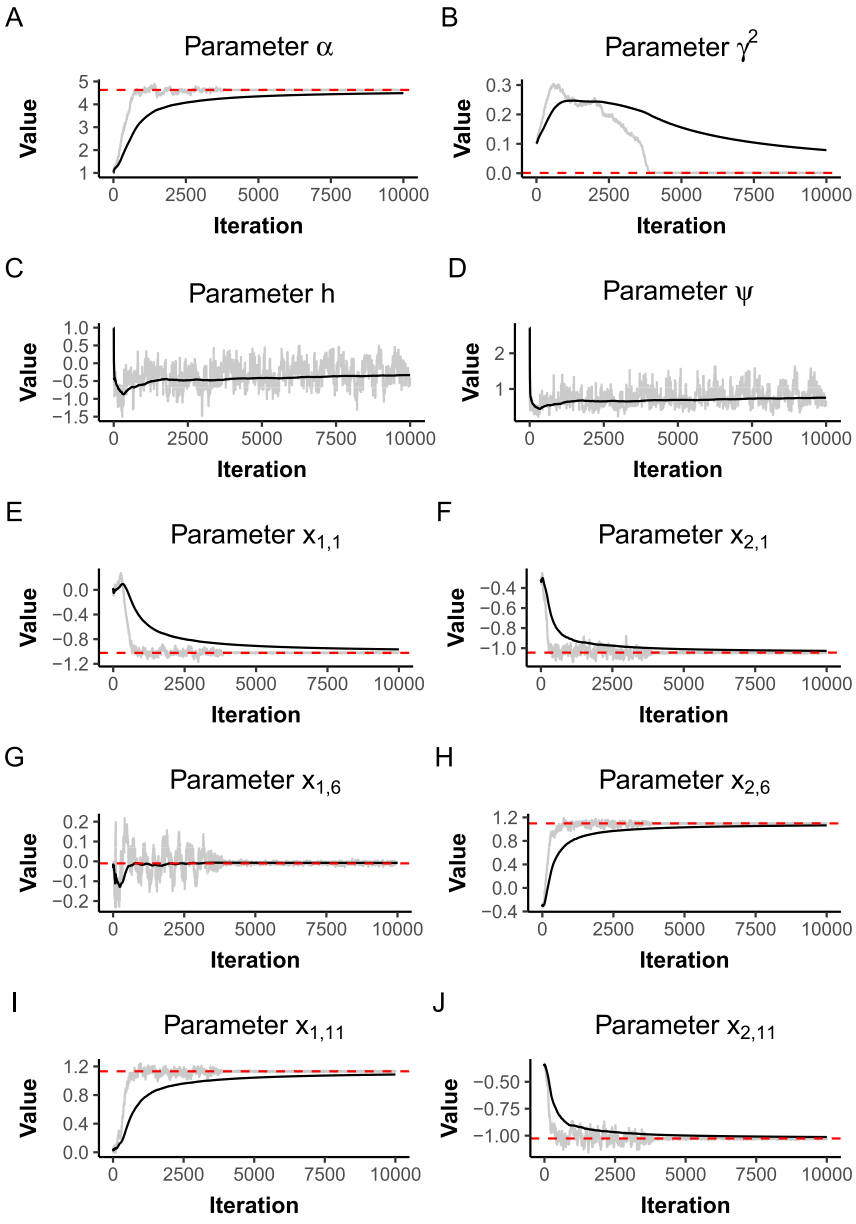

Figure 10: Simulation output. Trace plots for α_2 (Panel A), γ_2^2 (Panel B), h_2 (Panel C) and ψ_2 (Panel D), coordinates $\mathbf{x}_{1,2}$ (Panels E and F), coordinates $\mathbf{x}_{6,2}$ (Panels G and H), coordinates $\mathbf{x}_{11,2}$ (Panels I and J).

Appendix C: Further Details for the Empirical Application

C.1 Z-Transformations

The Fisher Z-transformation of a Pearson correlation coefficient C_{ijt} is defined as its inverse hyperbolic tangent (arctanh):

$$w_{ijt} = \frac{1}{2}\ln\left(\frac{1+C_{ijt}}{1-C_{ijt}}\right) = \operatorname{arctanh}(C_{ijt})$$

The transformation makes correlation coefficients approximately normally distributed with stable variance (Figure 11).

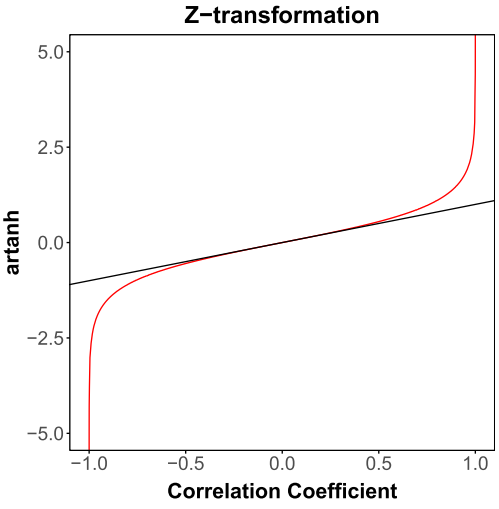

Figure 11: Z-transformation function.

C.2 Robustness Check: Longer Time Span

We assess whether our results in Section 4 are subject to substantial changes when a longer time window is considered. We increased the time span both for the DAX40 (from 2000 to 2022) and for the S&P100 (1990–2022). As constituents enter and exit from the two indices, the choice of the starting date is a trade-off between the number of constituents considered and the length of the time span. The number of constituents that were kept in the two indices for the whole length of the time spans is 25 for the DAX40 and 53 for the S&P100. Figure 12 displays the time series of volatility, average cross-correlation, and the average number of clusters for the constituents of both the DAX40 (left panel) and the S&P100 (right panel). The average number of clusters represents the Bayesian estimate of the number of clusters estimated using a rolling version of our dynamic LS model. The correlation of volatility and average cross-correlation with the average number of clusters is respectively 0.546 and 0.542 for DAX40 while 0.306 and 0.247 for the S&P100. Overall, these results seem to speak in favor of the existence of a positive correlation between clustering and periods of financial turmoil.

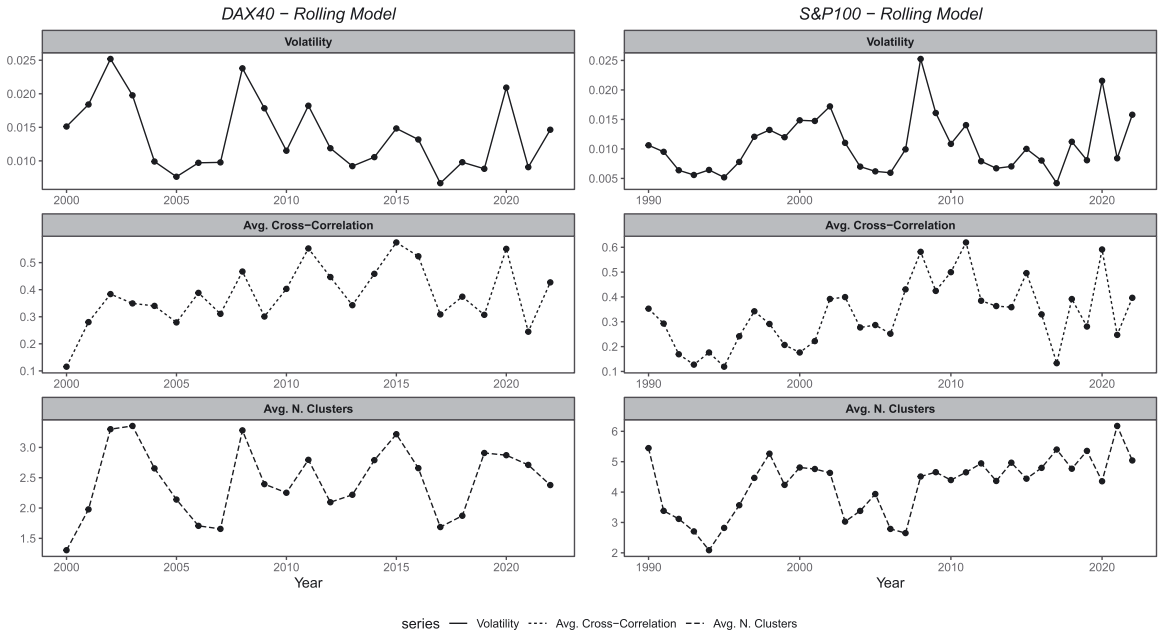

Figure 12: Robustness check: over-time volatility, average cross-correlation, and the average number of clusters for the constituents of both the DAX40 (left panel) and the S&P100 (right panel). The time series span from 2000 to 2022 in the case of the DAX40 and from 1990 to 2022 in the case of the S&P100. The average number of clusters represents the Bayesian estimate of the number of clusters from a rolling version of our dynamic latent-space model. The correlation of volatility and average cross-correlation with the average number of clusters is respectively 0.546 and 0.542 for DAX40 while 0.306 and 0.247 for the S&P100.

Appendix D: Code and Datasets

The authors published code and data associated with this article on Code Ocean, a computational reproducibility platform. We recommend Code Ocean to SNDE contributors who wish to share, discover, and run code in published research articles (see: https://doi.org/10.24433/CO.1842159.v1). Code and data are also available on GitHub (see: https://github.com/antonioperuzzi94/SNDE_LSM_AssetClustering).

References

Ahelegbey, D. F., L. Carvalho, and E. Kolaczyk. 2020. "A Bayesian Covariance Graph and Latent Position Model for Multivariate Financial Time Series." Available at SSRN 3090236.

Arbel, J., K. Mengersen, and J. Rousseau. 2016. "Bayesian Nonparametric Dependent Model for Partially Replicated Data: The Influence of Fuel Spills on Species Diversity." *The Annals of Applied Statistics* 10 (3): 1496—516.

Arnott, R. D. 1980. "Cluster Analysis and Stock Price Comovement." *Financial Analysts Journal* 36 (6): 56—62.

Barrat, A., M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. 2004. "The Architecture of Complex Weighted Networks." *Proceedings of the National Academy of Sciences* 101 (11): 3747—52.

Bassetti, F., R. Casarin, and F. Leisen. 2014. "Beta-Product Dependent Pitman-Yor Processes for Bayesian Inference." *Journal of Econometrics* 180 (1): 49—72.

Bassetti, F., R. Casarin, and F. Ravazzolo. 2018. "Bayesian Nonparametric Calibration and Combination of Predictive Distributions." *Journal of the American Statistical Association* 113 (522): 675—85.

Baştürk, N., L. Hoogerheide, and H. K. van Dijk. 2017. "Bayesian Analysis of Boundary and Near-Boundary Evidence in Econometric Models with Reduced Rank." *Bayesian Analysis* 12 (3): 879—917.

Billio, M., R. Casarin, and M. Iacopini. 2022a. "Bayesian Markov-Switching Tensor Regression for Time-Varying Networks." *Journal of the American Statistical Association*: 1—13. https://doi.org/10.1080/01621459.2022.2102502.

Billio, M., R. Casarin, M. Iacopini, and S. Kaufmann. 2022b. “Bayesian Dynamic Tensor Regression.” *Journal of Business & Economic Statistics* 41 (2): 1—11.
Billio, M., R. Casarin, and L. Rossini. 2019. “Bayesian Nonparametric Sparse VAR Models.” *Journal of Econometrics* 212 (1): 97—115.
Boccaletti, S., V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. 2006. “Complex Networks: Structure and Dynamics.” *Physics Reports* 424 (4—5): 175—308.
Bollobás, B. 1998. *Modern Graph Theory*, Vol. 184. New York: Springer Science & Business Media.
Bonanno, G., G. Caldarelli, F. Lillo, S. Micciche, N. Vandewalle, and R. N. Mantegna. 2004. “Networks of Equities in Financial Markets.” *The European Physical Journal B* 38 (2): 363—71.
Casarin, R., M. Costantini, and A. Osuntuyi. 2023. “Bayesian Nonparametric Panel Markov-Switching GARCH Models.” *Journal of Business & Economic Statistics*: 1—25. https://doi.org/10.1080/07350015.2023.2166049.
Casarin, R., D. Sartore, and M. Tronzano. 2018. “A Bayesian Markov-Switching Correlation Model for Contagion Analysis on Exchange Rate Markets.” *Journal of Business & Economic Statistics* 36 (1): 101—14.
Chesnay, F., and E. Jondeau. 2001. “Does Correlation between Stock Returns Really Increase during Turbulent Periods?” *Economic Notes* 30 (1): 53—80.
Christodoulakis, G. A. 2007. “Common Volatility and Correlation Clustering in Asset Returns.” *European Journal of Operational Research* 182 (3): 1263—84.
Conlon, T., H. J. Ruskin, and M. Crane. 2009. “Cross-Correlation Dynamics in Financial Time Series.” *Physica A: Statistical Mechanics and its Applications* 388 (5): 705—14.
Corsetti, G., M. Pericoli, and M. Sbracia. 2001. “Correlation Analysis of Financial Contagion: What One Should Know before Running a Test.” Technical report, Temi di discussione N. 408. Banca d’Italia.
Corsetti, G., M. Pericoli, and M. Sbracia. 2005. “Some Contagion, Some Interdependence: More Pitfalls in Tests of Financial Contagion.” *Journal of International Money and Finance* 24 (8): 1177—99.
Corsetti, G., M. Pericoli, and M. Sbracia. 2011. “Correlation Analysis of Financial Contagion.” In *Financial Contagion: The Viral Threat to the Wealth of Nations*, edited by R. W. Kolb, chapter 2, 11—20. Hoboken: John Wiley & Sons.
D’Angelo, S. 2018. “Latent Space Models for Multidimensional Network Data.” PhD thesis, 61—83.
Diestel, R. 2017. “Graph Theory.” In *Graph Theory*. Berlin: Springer.
Do, K.-A., P. Müller, and F. Tang. 2005. “A Bayesian Mixture Model for Differential Gene Expression.” *Journal of the Royal Statistical Society: Series C* 54 (3): 627—44.
Farrell, J. L. 1974. “Analyzing Covariation of Returns to Determine Homogeneous Stock Groupings.” *The Journal of Business* 47 (2): 186—207.
Fisher, M., and M. J. Jensen. 2022a. “Bayesian Nonparametric Learning of How Skill is Distributed Across the Mutual Fund Industry.” *Journal of Econometrics* 230 (1): 131—53.
Fisher, M., and M. J. Jensen. 2022b. “Bayesian Nonparametric Learning of How Skill is Distributed Across the Mutual Fund Industry.” *Journal of Econometrics* 230 (1): 131—53.
Forbes, K., and R. Rigobon. 2002. “No Contagion, Only Interdependence: Measuring Stock Market Co-movements.” *Journal of Finance* 57 (5): 2223—61.
Friel, N., R. Rastelli, J. Wyse, and A. E. Raftery. 2016. “Interlocking Directorates in Irish Companies Using a Latent Space Model for Bipartite Networks.” *Proceedings of the National Academy of Sciences* 113 (24): 6629—34.
Geweke, J. F. 1992. “Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments.” In *Bayesian Statistics*, edited by J. M. Bernado, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 169—93. Oxford: Clarendon Press.
Griffin, J., and M. Kalli. 2018. “Bayesian Nonparametric Vector Autoregressive Models.” *Journal of Econometrics* 203 (2): 267—82.
Griffin, J. E., and M. F. J. Steel. 2011. “Stick-Breaking Autoregressive Processes.” *Journal of Econometrics* 162 (2): 383—96.
Handcock, M. S., A. E. Raftery, and J. M. Tantrum. 2007. “Model-Based Clustering for Social Networks.” *Journal of the Royal Statistical Society: Series A* 170 (2): 301—54.
Heiberger, R. H. 2014. “Stock Network Stability in Times of Crisis.” *Physica A: Statistical Mechanics and its Applications* 393: 376—81.
Hirano, K. 2002. “Semiparametric Bayesian Inference in Autoregressive Panel Data Models.” *Econometrica* 70 (2): 781—99.
Hjort, N. L., C. Homes, P. Müller, and S. G. Walker. 2010. *Bayesian Nonparametrics*. New York: Cambridge University Press.
Hoff, P. 2021. “Additive and Multiplicative Effects Network Models.” *Statistical Science* 36 (1): 34—50.
Hoff, P. D., A. E. Raftery, and M. S. Handcock. 2002. “Latent Space Approaches to Social Network Analysis.” *Journal of the American Statistical Association* 97 (460): 1090—8.
Jensen, J. M., and M. J. Maheu. 2010. “Bayesian Semiparametric Stochastic Volatility Modeling.” *Journal of Econometrics* 157 (2): 306—16.
Kalli, M., J. E. Griffin, and S. G. Walker. 2011. “Slice Sampling Mixture Models.” *Statistics and Computing* 21 (1): 93—105.
Khandani, A. E., and A. W. Lo. 2007. “What Happened to the Quants in August 2007?” *Journal of Investment Management* 5 (4): 5—54.
Kim, D.-H., and H. Jeong. 2005. “Systematic Analysis of Group Identification in Stock Markets.” *Physical Review E* 72 (4): 046133.
King, B. F. 1966. “Market and Industry Factors in Stock Price Behavior.” *The Journal of Business* 39 (1): 139—90.
Kocheturov, A., M. Batsyn, and P. M. Pardalos. 2014. “Dynamics of Cluster Structures in a Financial Market Network.” *Physica A: Statistical Mechanics and Its Applications* 413: 523—33.
Müller, P., and R. Mitra. 2013. “Bayesian Nonparametric Inference — Why and How.” *Bayesian Analysis* 8 (2): 269—302.

Münnix, M. C., T. Shimada, R. Schäfer, F. Leyvraz, T. H. Seligman, T. Guhr, and H. E. Stanley. 2012. "Identifying States of a Financial Market." *Scientific Reports* 2 (1): 1—6.

Namaki, A., A. H. Shirazi, R. Raei, and G. Jafari. 2011. "Network Analysis of a Financial Market Based on Genuine Correlation and Threshold Method." *Physica A: Statistical Mechanics and its Applications* 390 (21—22): 3835—41.

Newman, M. 2018. *Networks*. Oxford: Oxford University Press.

Nie, C.-X. 2017. "Dynamics of Cluster Structure in Financial Correlation Matrix." *Chaos, Solitons & Fractals* 104: 835—40.

Nieto-Barajas, L. E., and F. A. Quintana. 2016. "A Bayesian Non-Parametric Dynamic AR Model for Multiple Time Series Analysis." *Journal of Time Series Analysis* 37 (5): 675—89.

Onnela, J.-P., A. Chakraborti, K. Kaski, J. Kertesz, and A. Kanto. 2003. "Asset Trees and Asset Graphs in Financial Markets." *Physica Scripta* 2003 (T106): 48.

Papastamoulis, P. 2014. "Handling the Label Switching Problem in Latent Class Models via the ECR Algorithm." *Communications in Statistics-Simulation and Computation* 43 (4): 913—27.

Preis, T., D. Y. Kenett, H. E. Stanley, D. Helbing, and E. Ben-Jacob. 2012. "Quantifying the Behavior of Stock Correlations under Market Stress." *Scientific Reports* 2 (1): 1—5.

Raftery, A. E. 2017. "Comment: Extending the Latent Position Model for Networks." *Journal of the American Statistical Association* 112 (520): 1531—4.

Rastelli, R., N. Friel, and A. E. Raftery. 2016. "Properties of Latent Variable Network Models." *Network Science* 4 (4): 407—32.

Sewell, D. K., and Y. Chen. 2016. "Latent Space Models for Dynamic Networks with Weighted Edges." *Social Networks* 44: 105—16.

Smith, A. L., D. M. Asta, and C. A. Calder. 2019. "The Geometry of Continuous Latent Space Models for Network Data." *Statistical Science* 34 (3): 428.

Taddy, M. A., and A. Kottas. 2009. "Markov Switching Dirichlet Process Mixture Regression." *Bayesian Analysis* 4 (4): 793—816.

Walker, S. G. 2007. "Sampling the Dirichlet Mixture Model with Slices." *Communications in Statistics-Simulation and Computation* 36 (1): 45—54.

Xu, Y., P. Müller, A. S. Wahed, and P. F. Thall. 2016. "Bayesian Nonparametric Estimation for Dynamic Treatment Regimes with Sequential Transition Times." *Journal of the American Statistical Association* 111 (515): 921—50.

Zhang, L., M. Guindani, F. Versace, J. M. Engelmann, and M. Vannucci. 2016. "A Spatiotemporal Nonparametric Bayesian Model of Multi-Subject fMRI Data." *The Annals of Applied Statistics* 10 (2): 638—66.

Zheng, Z., B. Podobnik, L. Feng, and B. Li. 2012. "Changes in Cross-Correlations as an Indicator for Systemic Risk." *Scientific Reports* 2 (1): 1—8.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/snde-2022-0111).